

Enough warheads, already

Its official: the reliability of existing US nuclear warheads makes their replacement unnecessary.

This week, UK prime minister Tony Blair declared his intention to replace the Trident submarine fleet that carries Britain's nuclear weapons. His declaration has sparked a debate in Britain about the role of nuclear weapons in the modern world — and about the obligation of states with nuclear weapons to move towards disarmament under the 1968 Nuclear Non-Proliferation Treaty (NPT).

But as Britain ponders its submarines, the United States is conducting a lower-profile but even more fundamental debate on whether to replace the nuclear warheads themselves. Over the past five years, some US weapons specialists have pushed a variety of new weapons concepts. But since the end of the cold war, making the case for brand new nuclear weapons has, thankfully, been a tough sell.

In Congress, liberals tend to argue that such weapons would defy the NPT. Conservatives see them as overly expensive. The military is happy with the nuclear weapons it already has, and the president, like his predecessors, wishes to continue a voluntary moratorium on nuclear testing.

But supporters of new and improved nuclear weapons are nothing if not tenacious. These days their main vehicle is a proposed project to build a Reliable Replacement Warhead (RRW). The argument for its pursuit goes, taking the words in reverse order, like this. It is a warhead, so it would require the nuclear-weapons labs to bring their collective knowledge to bear on its design and production. But it is merely a replacement, not a new design, so it won't violate international obligations, and won't, it is said, require full testing. And it is reliable — the clear implication being that existing warheads are not, or will not be in the near future.

It was this last argument that gave some urgency to the RRW programme. When the idea was conceived, it was estimated that the plutonium 'triggers' that fire the existing nuclear warheads might last only 40–60 years. As most of them are already 20 years old, it could

be argued, it was time for the laboratories to start work right away on finding a reliable replacement.

Unfortunately for advocates of the programme, however, it has emerged that existing warheads are more reliable than envisaged. Last week, the US National Nuclear Security Administration, which manages the nuclear-weapons programme, released unclassified summaries of two studies that found that the existing plutonium triggers will last for up to a century (see page 660). And that's in the worst case — the triggers may well last even longer.

Advocates of the replacement programme inside the Bush administration are apparently undaunted. Speaking through the Nuclear Weapons Council, which comprises the leaders of the agencies in the federal government that have an interest in these matters, they argued that the principal role of the RRW would, in fact, be to augment the security and safety of existing warheads. But these things are already more than adequately assured, through the decade-old stockpile stewardship programme.

Even before the new studies were released, the RRW seemed to be stuck in Congress. With Democrat gains in last month's mid-term elections, its prospects are bleaker still. Already, two senators, Jeff Bingaman (Democrat, New Mexico) and Dianne Feinstein (Democrat, California), have made their opposition to the programme clear. The political will to pursue it simply isn't there.

That's just as well. No convincing case has been made for the United States to replace its existing nuclear warheads. Their dreadful function is dependent on principles that were proven 50 years ago and aren't going to change any time soon. The government should give up its desire for new designs and settle on a size and shape for a research programme that fits the less extravagant needs of its existing nuclear-weapons stockpile — a stockpile that, with luck, will become obsolete before it needs to be replaced. ■

A fair share

The concept of sharing primary data is generating unnecessary angst in the psychology community.

In psychology there is little tradition of making the data on which researchers base their statistical analyses freely available to others after publication. This makes it difficult for anyone to independently reanalyse research results, and prevents small data sets from being combined for meta-analysis, or large ones mined for fresh insights or perspectives.

Psychologists need to rethink their reluctance to share data. Their discipline is 'softer' than some others: rarely do data on issues such as playground bullying or the usefulness of psychotherapy reveal really clear-cut answers. This makes the rigour with which the data

are handled fundamental to research outcomes — and increases the desirability of having them open to examination by peers.

The need for more data sharing has just been amply demonstrated by Jelte Wicherts, a psychologist specializing in research methods at the University of Amsterdam, who tried to check out the robustness of statistical analyses in papers published in top psychology journals.

He selected the November and December 2004 issues of four journals published by the American Psychological Association (APA), which requires its authors to agree to share their data with other researchers after publication. In June 2005, Wicherts wrote to each corresponding author requesting data, in full confidence, for simple reanalysis. Six months and several hundred e-mails later, he abandoned the mission, having received only a quarter of the data sets. He reported his failure in an APA journal in October (J. M. Wicherts *et al. Am. Psychol.* **61**, 726–728; 2006).

Researchers often have valid reasons for constraining access to their



raw data, such as the privacy of research subjects. But data from most studies based on confidential information can be coded in a way that will guarantee their subjects' anonymity. The few cases where this is not possible can be exempted from the move towards data sharing.

A second factor deterring openness is a natural desire to retain exclusive access to data that took years of care and attention to collect. Like many researchers in other disciplines, psychologists fear that if different analytical approaches are brought to bear on their data, different conclusions could be drawn, casting doubt on their competence — or even their integrity. But in most cases, if data have been collected, selected and analysed correctly, researchers have little to fear in this regard, and the resulting discussion is likely to prove enlightening for the field as a whole.

An associated concern is that data could be wilfully misinterpreted by anyone with a political agenda. But this should not prevent the sharing of data sets: false interpretations of the data will fail to find any foothold in the community as whole.

A less frequently articulated reason for resistance to data sharing is the fact that some researchers are simply unable or unwilling to record and present their data in an unambiguous, reader friendly and archivable form.

The APA's editors and publishers are now planning their response to Wicherts' report. One result should be the acceleration of moves, already under discussion, to require the deposition of data as supplementary electronic material in APA databases. Where the APA leads, other psychology journals are likely to follow.

Granting bodies must also play a part. In 2003, the US National Institutes of Health introduced rules requiring the public sharing of data in psychology studies for grants exceeding \$500,000, allowing exemptions where confidentiality issues cannot be circumvented. Other agencies should follow suit. And university departments need to do more to teach the basics of note-keeping and data presentation, to prepare their students for an era in which data sharing will increasingly become the norm. ■

Green shoots of growth

Energy from biomass is an idea whose time has returned.

Until the twentieth century, biomass was humanity's principal source of energy, heating our stoves and feeding our draught animals. Even today, roughly 10% of all our energy comes from biomass — far more than from any other renewable energy source or, for that matter, from nuclear fission.

But this use of biomass for energy supply is accompanied by many challenges (see page 669). For one thing, it is often not all that renewable — the biomass sources that provide firewood to the world's poor, for example, are not being replanted. For another, it is very inefficient: gathering firewood takes a long time. The history of the past couple of centuries has been in large part one of people moving away from biomass as soon as they can afford to do so.

Three recent developments have spurred renewed interest in biomass, however. One is the need to reduce greenhouse-gas emissions. The requirement for other external energy inputs during biomass processing means that it often involves some net carbon emissions — but the amount of carbon dioxide given off by burning biomass is the same as that taken from the atmosphere by photosynthesis in the first place. If biomass projects could sequester carbon, either by enriching the soil beneath plantations or by storing any carbon dioxide produced in combustion, they could even be carbon negative — a unique selling point for this energy source.

The other two developments are the upward movement in the prices of oil and natural gas, and the related revival of concerns about the security of their supply. Most nations are seeking home-based energy sources that do not rely on political stability in the Middle East or Russia.

It seems unlikely that these factors will provide sufficient impetus to propel biomass energy to the very front rank of possible alternatives to fossil fuels. But biomass clearly has a potential role as part

of a portfolio of energy sources for the twenty-first century.

If that role is to be fulfilled, two things need to happen. Nations have to build regulatory mechanisms that recognize the carbon benefits of technologies such as biomass — through emissions pricing, a carbon tax or a combination of the two. And intensive research needs to be conducted into both the efficient production of biomass and its conversion into useable energy.

One focal point for such research should be finding ways to grow biomass quickly and in an easily processed form while minimizing external inputs, such as fertilizer and pesticides. Another is the systems engineering of farms and ecosystems, finding ways to fit biomass projects into and around present land use and possible changes in farming practice.

A major attraction of biomass is that it is likely to benefit poorer countries, which tend to be in tropical regions where plants grow quickly. There is plenty of scope for more collaboration between developing countries on biomass research and development, both to meet local needs and for export.

But this requires consideration of the local and global ecological impact of biomass expansion. Vast tropical monocultures eating away at primary forests — as exemplified by the production of palm oil in Indonesia — will benefit no one, except those who profit from selling the fuel. In effect, such approaches take green subsidies from richer countries, and use them to despoil the tropics.

Similar problems afflict existing biomass programmes in the United States, where ethanol refineries often burn fossil fuel and are reliant on subsidized corn monoculture. More innovative approaches would include firing the refineries with agricultural waste, and feeding them with plants of many different species. Biomass energy should be developed energetically, but within the context of appropriate environmental policies, and using approaches that are both sustainable and cost-effective. ■

"Biomass is likely to benefit poorer countries, which tend to be in tropical areas where plants grow quickly."

RESEARCH HIGHLIGHTS

Blockbuster film

Phys. Rev. E **74**, 051602 (2006)

Could you make a soap film twice the size of Tiananmen Square without it popping? In space it might be possible, say Rui Zheng and Thomas Witten of the University of Chicago, Illinois.

They suggest using, instead of soapy water, a commercial pump oil stabilized by surfactants, which would evaporate slowly enough to create a film that is 1 kilometre across by 1 micrometre thick and stable for at least a year. In Earth orbit, the film would absorb enough sunlight not to freeze. And with the right surfactant, the film shouldn't rupture spontaneously; meteoroids that punch holes, however, are more problematic.

Why go to all the bother? The film would be a great laboratory for investigating two-dimensional flow and turbulence, the authors argue — important in geophysics and astrophysics but hard to study in the lab.



N. JORGENSEN/REX FEATURES

CELL BIOLOGY

DNA is pulled into line

Genes Dev. **20**, 3269–3282 (2006)

When cells divide, they replicate their DNA so that each daughter cell can receive a full chromosomal complement. Eukaryotes use specialized motor proteins to pull chromosomes into each new cell, but how bacteria divvy up genetic material has remained a mystery.

Michael Fogel and Matthew Waldor of Tufts University in Boston, Massachusetts, now show how the cholera-causing bacterium, *Vibrio cholerae*, distributes its chromosomes when dividing. A protein called ParAI stretches across the cell, grabs a chromosome, and retracts, pulling the DNA along with it.

The finding provides the first example of pulling forces that distribute bacterial DNA, and suggests a possible evolutionary link between chromosome segregation in bacteria and in eukaryotes.

NANOTECHNOLOGY

Spare the rod

J. Am. Chem. Soc. doi:10.1021/ja064535p (2006)

Nanorods of an organic material increase in length by an average of 15% when illuminated with ultraviolet light, report Christopher Bardeen and his colleagues at the University of California, Riverside. Surprisingly, randomly shaped crystals of the same material shatter under identical treatment.

Ultraviolet light induces pairs of molecules in the nanorods to react with each other, forming a product that takes up more molecular space than the starting material. The resulting crystal strain is more easily released at the surface of the rod than within it. The high surface-to-volume ratio of nanorods, the authors propose, allows strain to be released non-destructively by the extending rods.

This effect could eventually be harnessed to trigger nanoscale motion from light energy, the researchers suggest.

IMMUNOLOGY

Body invaders

J. Clin. Invest. **116**, 3258–3265 (2006)

In the disease type-1 diabetes, the body's immune system attacks many different proteins. Previous research had shown that the first victim is proinsulin. But until now, it was not clear whether immune responses spread from there to other proteins, or whether those others were attacked simultaneously.

Thomas Kay from St Vincent's Institute in Victoria, Australia, and his colleagues used genetically modified mice to study two proteins — proinsulin, and another called IGRP — that trigger inflammatory responses. Mice engineered to not respond to proinsulin also had no response to IGRP, and did not get sick. But those designed not to respond to

IGRP did recognize proinsulin and developed diabetes — suggesting that the response to proinsulin is 'upstream' of that to IGRP.

Proinsulin triggers the first and key response, the authors suggest, and thus should be the main target for diabetes treatments early in the course of the disease.

SUPRAMOLECULAR CHEMISTRY

Tying the knot

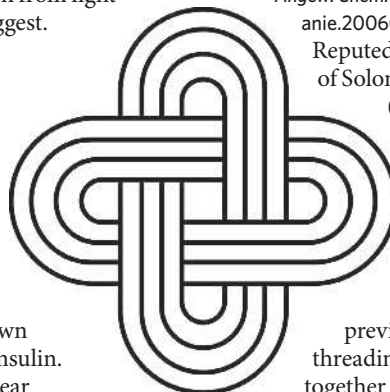
Angew. Chem. Int. Edn doi:10.1002/anie.200603521 (2006)

Reputed to encode all the wisdom of Solomon, King Solomon's Knot (pictured left) is found in sacred art from Africa to Celtic Britain. An interweaving of two chain links, it is not strictly speaking a knot, but a link.

Chemists have previously figured out ways of threading two molecular hoops together to make Solomon's link, but the method now described by Fraser

Stoddart of the University of California, Los Angeles, and his co-workers has a new twist. They assemble the link from 12 distinct components, using metal ions to template the threading of the organic components of the two rings.

The Solomon link is just one of the possible interwoven products, but its formation can be amplified during crystallization, promoting this molecule over the others.



PHYSIOLOGY

To flee or not to flee

Proc. R. Soc. Lond. B. doi:10.1098/rspb.2006.3755 (2006)

If not exposed to predators regularly, animals seem to lose the physiological triggers that could save them from being eaten, a new study suggests.

For millions of years, until dogs and cats showed up a century and a half ago, marine iguanas (pictured right) lived in the Galapagos Islands without major land predators. In an experiment, researchers chased iguanas that still live in isolated areas; the animals, they found, did not generate the stress hormone corticosterone. But iguanas living near a town, and presumably more used to being attacked by predators, did produce the hormone.

The iguanas were fast learners, however. Even the naive animals quickly started producing the hormone after being chased a couple of times, says a team led by Thomas Rödl of Princeton University, New Jersey.



When joining sugars together, adding an extra molecular group to one of them allows enzymes called glycosyltransferases to link them up in different configurations. This relatively simple strategy can produce a range of strikingly different products. Synthetic chemists, take note.

MEDICINE

A one-two punch for cancer

Science **314**, 1308–1311 (2006)

Adding a genetically modified bacterium to chemotherapy drugs fights cancer better than the drugs alone, new work suggests.

Getting cancer treatments to kill cancerous, but not healthy, cells is a major problem for drug delivery. Now, a team led by Shibin Zhou and Bert Vogelstein at the Johns Hopkins Kimmel Comprehensive Cancer Center in Baltimore, Maryland, says that adding the bacterium *Clostridium novyi-NT* can help.

The combination treatment caused both large and small tumours in mice to shrink. The bacterium may draw the drug more selectively to the cancer cells, helping it to better penetrate and destroy the tumour.

GEOLOGY

Drilling deep

Geophys. Res. Lett. **33**, L21314 (2006)

The first-ever deep-drilling programme into California's infamous San Andreas Fault is starting to yield information about the mineralogy of the rocks there.

John Solum of the US Geological Survey in Menlo Park, California, and his colleagues have identified several distinct mineral assemblages, down to nearly the bottom of their 4,000-metre-deep hole.

In particular, the mineral serpentine — sometimes seen along the San Andreas at the surface — appeared in the layer where the fault was actively deforming the rocks. The presence of serpentine, some think, could help explain certain details of how the fault behaves.

Coring at the San Andreas Fault Observatory at Depth, a US government project, begins again next summer.

PHYSICS

Liquid-controlled light

Appl. Phys. Lett. **89**, 211117 (2006)

The passage of light through a layer of silicon can be controlled using liquid droplets less than 1 femtolitre in volume.

Francesca Intonti of the European Laboratory for Non-Linear Spectroscopy in Firenze, Italy, and her colleagues placed drops of water into selected pores in a silicon wafer. Light moving through the silicon was then constrained to the S-shaped path marked out by the drops.

The technique could provide a new method for constructing analogues of electronic semiconductor devices, such as optical switches.

CHEMISTRY

Same key, different lock

Nature Chem. Biol. **2**, 724–728 (2006)

Enzymes are so good at chemical reactions that they are sometimes too good to be useful when making new compounds. The enzymes are too specific, and always carry out the same reaction. One way around this is to engineer or evolve the enzymes to do what's wanted of them. Another solution, highlighted by Stephen Withers at the University of British Columbia in Vancouver, Canada, and his group, is to instead tweak the substrate, or landing pad for the enzyme, through a temporary modification.

JOURNAL CLUB

Lewis E. Kay
University of Toronto, Canada

The molecular dance of a protein allows a chemist's secret wish to come true.

One fascinating aspect of molecular function is the way information propagates between parts of a molecule that can be many tens of angstroms apart.

Our understanding of how proteins do this, a process termed allostery, emerged from Max

Perutz's pioneering studies of oxygen-carrying haemoglobin. Three-dimensional images show that when a ligand binds to part of the molecule, a discrete set of structural changes take place at distinct sites. This, in turn, influences the ease with which subsequent ligands bind.

Nature has chosen this model in designing many allosteric proteins. However, as a practising nuclear magnetic resonance (NMR) spectroscopist with a strong interest in protein dynamics, I was secretly hoping she might design

proteins in which information is communicated through changes in the dynamics between distal sites, with little or no change in overall structure. Moreover, I was rooting for NMR to play a major role in characterizing such a system.

How exciting it was, therefore, to read that Charalampos Kalodimos and his co-workers recently found such a case by studying the motional properties of a protein in different ligated states (N. Popovych *et al.* *Nature Struct. Mol. Biol.* **13**, 831; 2006). Using NMR spectroscopy, the team quantified

protein dynamics for a wide range of timescales. Remarkably, ligand binding at one site is linked to changes in motion far removed, over the complete set of timescales, while a corresponding propagation of structural changes does not occur.

The work of Popovych *et al.* provides a striking example of the importance of protein dynamics to information transfer. I eagerly await the discovery of more molecular dances and of how they, too, will relate to biological function.

NEWS

Journal reveals plans to fight fraud

WASHINGTON DC

In a notable move aimed at curbing fraud in scientific publications, the journal *Science* said last week that it will probably begin targeting certain “high-risk” papers for extra scrutiny.

The move comes in response to a report from an external committee convened by the journal to assess its handling of the papers behind the Woo Suk Hwang fraud scandal. And it turns on its head — for a handful of papers at least — the traditional presumption that manuscripts submitted to a journal are researched and written honestly.

“Until now, it has been assumed as a default that scientists are honest. The burden of proof is to show that they are not. Now, at least for a select number of papers where the risk factor is high, there is a new burden, to show that these papers are honest,” says Sheldon Krinsky, a bioethicist at Tufts University in Medford, Massachusetts.

Hwang, a South Korean researcher working at Seoul National University, published high-profile papers in *Science*^{1,2} in 2004 and 2005 that claimed to have generated embryonic stem cells by somatic-cell nuclear transfer. This is a key step to generating replacement tissues from a patient’s own cells. Both papers turned out

to have been fabricated, and *Science* retracted them in January³.

The review committee, headed by John Brauman, a chemist at Stanford University in California, released its report into *Science*’s conduct last week. Although *Science* had followed high-standard editorial procedures “with exceptional care” in the Hwang case, Brauman says, “we suggested that the journal institute a policy we describe as risk assessment” in an effort to clamp down on fraud.

Writing in an editorial⁴, *Science*’s editor-in-chief Donald Kennedy said that the journal is now developing criteria for deciding which papers deserve particularly careful scrutiny. “Papers that are of substantial public interest, present results that are unexpected and/or counterintuitive, or touch on areas of high political controversy may fall into this category,” he wrote.

Such papers, perhaps ten a year, would receive “special attention” that could include greater requirements for including primary data and more intensive evaluation of digital images. The journal would also demand explicit descriptions of each author’s contribution to a paper.

Even as Kennedy announced the plans, *Science* reported doubts about the results of

another high-profile paper it recently published in a controversial area of developmental biology. Pending the outcome of an investigation by the University of Missouri, Columbia, this will also probably be retracted.

In a paper⁵ that sparked debate from the moment it was published, R. Michael Roberts and colleagues, researchers at the university, claimed that mouse embryonic cells have distinct developmental fates from the first cell division onwards. This flies in the face of the broadly held view that in mammals such cells can still develop into any cell in the body. The paper was published on 17 February this year, and the university launched an investigation in April that is still continuing.

Kennedy cautioned in a press briefing that the social costs associated with loss of trust among scientists might be greater than those of the occasional retraction. But he said he would collaborate with *Nature* and other journals to draw up a common set of standards aimed at deterring fraud. Philip Campbell, *Nature*’s editor-in-chief, declined to comment in detail on the committee’s findings, but said: “We at *Nature* welcome the external review conducted by *Science* and are considering its recommendations.”

Molecular HIV evidence backs accused medics

International experts in DNA forensics say that a paper published online by *Nature* this week provides a firm alibi for the six medical workers facing the death penalty in Libya. The workers have been charged with deliberately infecting more than 400 children with HIV in 1998.

In the study, an international team led by researchers from Oxford and Rome used the genetic sequences of the viruses isolated from the patients to reconstruct the exact phylogeny, or ‘family tree’, of the outbreak. Analysing the mutations that accumulated over time allowed the researchers to work out when different outbreaks occurred. They showed that the strain of HIV with which the children had been infected was already present and spreading locally in the

mid-1990s, long before the medics arrived in Libya in 1998.

The trial of the six medical workers ended in Tripoli on 4 November, and a verdict is expected on 19 December. Despite mounting international pressure to free them, defence lawyers are pessimistic about the outcome, and *Nature* has fast-tracked publication to make this new evidence available before the verdict (see *Nature* 444, doi:10.1038/nature444836a; 2006).

There was already a body of scientific evidence indicating that the outbreak was caused not by deliberate transmission, but by poor hygiene at the Al-Fateh hospital in Benghazi, where the outbreak took place (see *Nature* 443, 888–889; 2006). Analysis of hospital records suggested that the outbreak began before the medics arrived. And



Six medical workers are accused of deliberately infecting children with HIV.

almost half of the HIV-infected children were also infected with hepatitis B or C, pointing to poor hospital practices as the cause.

The new results provide independent genetic confirmation of these findings. As well as showing that the outbreaks

predated the medics’ arrival, the study suggests that the HIV strain is related most closely to strains from West Africa, suggesting a natural introduction, probably via the many migrant workers in Libya, says co-author Tulio de Oliveira of the University of Oxford, UK.



AIDS MEDICS IN LIBYA
Find all the news on this ongoing case online.
www.nature.com/nature/focus/aidsmedicslibya



High-impact papers, such as those submitted by Woo Suk Hwang, would be subjected to extra scrutiny.

But many commentators seem to feel that the recommendations don't go far enough. "There should be very strict criteria not only for high-impact or controversial papers but for everybody," says Thereza Imanishi-Kari,

an immunogeneticist at Tufts. Imanishi-Kari's career stalled when her work became the subject of a high-profile US government investigation in the 1990s. She was exonerated.

Krimsky also favours increased scrutiny

of papers: "What could the negative impacts be? A less trusting community of scientists? A new McCarthyism? I don't think that's likely." He adds that a further criterion should trigger extra scrutiny of papers: where an author or authors have an unusually high degree of commercial interest in the results.

Meanwhile, Arthur Caplan, director of the Center for Bioethics at the University of Pennsylvania in Philadelphia, says he would like to see the criteria expanded to include papers in areas with a history of fraud and those in highly specialized areas where relatively few other scientists would be equipped to detect such problems. He adds: "I think you have to be blunt and say that papers coming from certain countries also raise red flags. China has not yet been convincing and South Korea has problems."

But Kennedy said in the press briefing that he would guard against any targeting of foreign scientists. "We don't want to engage in profiling. It would really be unfair if we started looking extra hard at some of the papers from emerging scientific powers in countries such as South Korea."

Meredith Wadman

1. Hwang, W. S. *et al.* *Science* **303**, 1669-1674 (2004).
2. Hwang, W. S. *et al.* *Science* **308**, 1777-1783 (2005).
3. Cyranoski, D. *Nature* **439**, 122-123 (2006).
4. Kennedy, D. *Science* **314**, 1353 (2006).
5. Deb, K., Sivaguru, M., Yong, H. Y. & Roberts, R. M. *Science* **311**, 992-996 (2006).

S.-J. CHUNG/GETTY/NEWS.COM

Other phylogenetic analyses of HIV have been used in court cases involving allegations of HIV infection. The first was in 1991, when a Florida dentist was shown to have contaminated his patients. The technique has since been accepted as evidence in dozens of cases involving rape, hospital infection transmissions and people with HIV knowingly exposing others in Sweden, France and elsewhere.

Thomas Leitner of Los Alamos National Laboratory in New Mexico has provided forensic HIV evidence in more than 30 such cases over the past 15 years. He describes the de Oliveira paper as "compelling evidence that the outbreak had started before the accused could have started it", a view shared by every expert that *Nature* contacted (see 'Expert opinion').

Leitner points out that calculating evolutionary timescales is tricky, but that because HIV has such a fast mutation rate, even recent events can be pinpointed quite accurately. "De Oliveira *et al.* have tested and evaluated the clock and its uncertainty using several methods," he says. "I find their analysis well done and timely, and hope it will affect the judgement in the Libyan court." ■ Declan Butler

Expert opinion

The following scientists have all previously testified in court cases involving HIV molecular evidence. They assess the new data.

"This study is an impressive statistical analysis. It shows clearly that the hypothesis of deliberately injecting children with HIV in 1998 should be rejected."

Philippe Lemey, expert on HIV evolution, Rega Institute for Medical Research, Belgium.

"This is exactly the kind of objective phylogenetic analysis needed in this case. The results clearly show that the health workers were not responsible for the introduction of these HIV strains."

David Hillis, expert on viral phylogenies, the University of Texas, Austin.

"This kind of analysis has been approved by courts around the world. This is a case of [hospital] infection with multiple, independent sources, a pattern most easily explained by sloppy or inappropriate practices at the hospital."

Fernando González Candela, evolutionary geneticist, the University of Valencia, Spain.

"The existing epidemiological data are already enough to demonstrate that the accused medical staff cannot be the source of the contamination. De Oliveira's analysis is completely independent,

and yields the same conclusion. The court cannot pretend to be impartial if it refuses to hear any competent scientist from abroad."

Michel Milinkovitch, evolutionary geneticist, the Free University of Brussels, Belgium.

"They have used state-of-the-art methods to estimate divergence and dates of events in this outbreak. The analysis shows compelling evidence that the outbreak had started before the accused could have started it."

Thomas Leitner, expert in HIV evolution, Los Alamos National Laboratory, New Mexico.

US backs revamp of nuclear warheads

The US government is pushing ahead with a plan to overhaul its nuclear stockpile, despite a scientific review showing that existing warheads will last at least another half-century.

On 1 December, the National Nuclear Security Administration — the agency that oversees the US nuclear stockpile — announced that it will pursue plans to develop a new Reliable Replacement Warhead (RRW), which it claims will be safer and more robust than current designs. But the announcement came just two days after the release of research showing that the plutonium in existing warheads has a shelf-life of at least a century. Most warheads in the US stockpile are only about 20 years old.

Critics say the decision to proceed is an example of politics trumping science. "It is clear that the present stockpile is going to be reliable beyond our lifetime," says Jay Coghlan, executive director of Nuclear Watch, a watchdog group in Santa Fe, New Mexico. "It's not the science that rules, it's the special interests."

But officials maintain that the programme was never solely about replacing old warheads. "I believe the reasons for the RRW are in some ways independent of plutonium ageing," says Charles McMillan, associate director for weapons physics at Los Alamos National Laboratory in New Mexico. "RRW enables us to bring new technologies to the table."

Since 2004, the RRW concept has been

gathering momentum — both at the nation's nuclear weapons labs and in Congress (see *Nature* **442**, 18–21; 2006). The idea is to build a new generation of robust warheads to replace existing designs. They would be larger than existing warheads, but would have a greater margin of error, be easier to manufacture and include more safeguards.

The proposed designs have mainly focused on making changes to the plutonium triggers, or pits, of the current warheads. Some argue that adding more plutonium or changing the way they are cast would result in a more stable, dependable trigger that could sit on the shelf for years without testing.

Officials have pushed for the programme to proceed quickly because of concerns that the ageing triggers on the current warheads might become unreliable within a few decades. The chief worry was that the constant stream of radiation that comes from the plutonium itself could create cavities in its crystalline structure, causing the weapon to fail.

But a trawl of old nuclear test data and a battery of lab tests have shown this isn't the case. Some of the 1,054 nuclear tests carried out by the United States before 1992 were done with ageing weapons, and the data helped scientists

understand how older plutonium triggers behave. Meanwhile, researchers also discovered that artificially aged plutonium heals itself by shifting new atoms back into its crystalline lattice. The current generation of weapons can therefore last for at least 85–100 years.

These findings are likely to have implications elsewhere, especially in Britain, which is thought to have warheads similar to the US design. Other nations such as France use plutonium triggers and have stockpile stewardship programmes.

The findings were "a real surprise," says

Raymond Jeanloz, a geologist at the University of California, Berkeley, and a member of the JASONs — an independent scientific group that advises the US government on security issues and that reviewed the work. "The labs have done

an outstanding job."

The revised lifetimes call into question the need for an RRW programme, say many critics. RRW was sold to congressional supporters on the basis that the ageing stockpile would soon have to be replaced anyway. But as it will be in good condition for the conceivable future, such a plan now seems unnecessary, says Christopher Paine, a nuclear analyst at the Natural Resources Defense Council, an environmen-

"The arguments with which replacement warheads won congressional support have fallen apart."

Bush faces rough ride over climate change

With President George W. Bush and his administration continuing to avoid the issue, the courts and Congress are poised to shape climate-change policy in the United States.

Last week, the Supreme Court heard a high-profile case in which Massachusetts, along with numerous other states, cities and environmental groups, argued that the Environmental Protection Agency (EPA) should be forced to regulate greenhouse-gas emissions from cars and trucks as a pollutant. Massachusetts itself stands to lose 300 kilometres of its coastline as a result of the rise in sea level as the planet warms.

The suit can be read as an

expression of many states' impatience with the federal government's inaction on climate change (see *Nature* **443**, 486–487; 2006). Meanwhile, staff in the offices of Democratic politicians are sharpening their pencils for

January, when the more liberal party takes over Congress, and supporters hope that climate-change bills will fly.

The verdict will be announced by the Supreme Court by the end of its term in June. Although the EPA trotted out the same old administration line that the "scientific uncertainty" was too great to act upon, the discussions

during the 29 November hearing hint that the case might be decided on the technical grounds of standing — Massachusetts could sue the agency only if the harm suffered will be redressed if the EPA does act.

The problem is that climate change is such a large and global problem that the emissions in question — from the tailpipes of US vehicles — may not make a huge difference by

themselves. Massachusetts will probably still lose a considerable chunk of coastline.

Beneath the issue of standing, there seems to be a schism of

opinion on more ideological grounds, with the more conservative judges looking likely to support the EPA, and the more liberal judges lining up with the Massachusetts group. With four justices on each side, Justice Anthony Kennedy, a noted moderate, may have the casting vote.

But many observers say that the decision itself will be less important than the buzz around the case. "I think no matter which way it breaks, it will put more pressure on Congress," says Andrew Aulisi of the World Resources Institute, an environmental think-tank in Washington DC. If the Supreme Court sides with the EPA, he says, "everybody is going to throw up

"I think no matter which way the law suit breaks, it will put more pressure on Congress."



P. SHAMBROOM/NUKEPHOTO.COM

Research suggests that the present crop of US nuclear warheads should last for at least 80 years.

tal group based in New York. "The arguments with which they won congressional support have fallen apart."

McMillan counters that the RRW has always been about more than just replacing older warheads. "RRW will bring to the stockpile the most modern technologies for safety and security," he says. In addition, he says, the process of

developing and producing the RRW will help transform the weapons complex into a smaller, more responsive one.

If nothing else, the extended shelf-life of the current warheads should allow more time for debate.

Geoff Brumfiel

See Editorial, page 653.

their hands and say that the judiciary isn't doing anything and Congress needs to step in".

When the Democrats take control of Congress in January, the prediction for the first months is a steady diet of hearings, with bills taking a little longer. No one is quite sure if the votes are there for a tough bill on climate change. "What the elections did, to a large extent, was replace Republican moderates with Democratic moderates," says Manik Roy, director of congressional affairs at the Pew Center on Global Climate Change. "You don't have many more votes. What has changed is who controls the agenda."

Senator Barbara Boxer (Democrat, California), who will

head the Senate Committee on Environment and Public Works, has promised lots of talk and action on climate change, and on 15 November sent a letter with two other heads of related committees to President Bush, asking for a commitment to "pass meaningful climate change legislation in 2007". Republican Senator (and probable presidential candidate) from Arizona, John McCain, will also undoubtedly reintroduce the McCain-Lieberman Climate Change Act, a cap-and-trade bill, for its third outing.

Frank Maisano, a spokesman for Bracewell & Giuliani, a law firm representing oil and gas industries, cautions that advocates for climate-change

regulation have been overtaken by unrealistic exuberance. The complexity of the issue will push substantive action way past the verdict, he says. "It is not something that is going to be slam-dunked in 12 months."

But even if the rounds of hearings seem to produce nothing but hot air, Roy points out that they will at least educate members of Congress, where, for example, Senator James Inhofe (Republican, Oklahoma) has been holding forth on his view of climate change as some sort of conspiracy theory. "I would not in any way consider it a delay tactic if Congress spends a year holding hearings on this issue," Roy says.

Emma Marris

ON THE RECORD

Justice Antonin Scalia

(pictured):

"Your assertion is that after the pollutant leaves the air and goes up into the stratosphere it is contributing to global warming."



B. CHILD/AP

James Milkey:

"Respectfully, Your Honour, it is not the stratosphere. It's the troposphere."

Justice Scalia:

"Troposphere, whatever. I told you before I'm not a scientist. That's why I don't want to have to deal with global warming."

The US Supreme Court tackles climate change (see left).

OVERHYPED

Radioactive products

After the news that Russian ex-spy Alexander Litvinenko was poisoned with polonium-210, a number of blogs and news stories sounded shrill warnings about companies selling polonium-210 over the Internet.

One such company, United Nuclear in Sandia Park, New Mexico, posted a notice on its website explaining that the amounts of polonium-210 it sells are microscopic. The company estimates it would take about 15,000 of its polonium-210 sources to poison someone — at a cost of \$1 million.

"An order for 15,000 sources would look a tad suspicious," the company points out, "considering we sell about one or two sources every three months."



MEDICAL RESEARCH

A review of the top medical news of 2006.

www.nature.com/news/channels/medicalresearch.html

South Africa takes steps to tackle HIV

CAPE TOWN

At last, the South African government seems to have agreed to tackle the AIDS pandemic in a whole-hearted manner. Admitting that “we all know that HIV and AIDS are among us”, deputy President Phumzile Mlambo-Ngcuka appealed to South Africans on 1 December, World AIDS Day, to “join hands to save our people, in the same way that we joined hands to fight and defeat apartheid”.

At a rally in Nelspruit, the capital of South Africa's Mpumalanga province, Mlambo-Ngcuka released a draft of the government's plan. The document contains only two targets: to halve the rate of infection, and to make antiretroviral treatment available to 80% of those who require it, both by the end of 2011.

Mlambo-Ngcuka also declared her intention to restructure the South African National AIDS Council to include representatives from the private sector, trade unions and non-governmental organizations. No information is available on how the targets will be met — the government plans to finalize the details by March. But the developments have been greeted with cautious optimism by AIDS activists and practitioners.

“If enough funds are provided, and if human resources problems — particularly relating to nurses — can be resolved, treatment targets are achievable,” says economist Nicoli Nattrass, director of the AIDS and Society research unit at the University of Cape Town.

South Africa has previously taken an ambivalent approach to HIV. President Thabo Mbeki



South Africans face a brighter future now their government is set for serious action on HIV.

has maintained a stony silence on the issue, and with no coordinated national strategy to provide antiretroviral drugs to patients, access to treatment has been patchy at best.

Many see the World AIDS Congress, held in Toronto, Canada, in August, as a turning point. The government was embarrassed by health minister Manto Tshabalala-Msimang, who used it to extol the virtues of garlic, lemon and beetroot in the fight against AIDS. Mlambo-Ngcuka, who many see as a possible successor to Mbeki, quickly assumed responsibility for government policy on the issue.

She faces a tough task. A report released last week by the Actuarial Society of South Africa (ASSA) stated that 1.8 million people have died

of AIDS in South Africa since the early 1990s. Between 800,000 and 1 million people need antiretroviral treatment but only a third of them — and a tenth of children — are getting it.

Slowing down rates of infection might be even more difficult. The ASSA estimates that half a million people became infected this year, down from 650,000 in 1998.

But the signs are promising. In 2005, 1.7 million people used government counselling and testing services, and this year the number of people treated for HIV nearly doubled. Last week the deputy health minister, Nozizwe Madlala-Routledge, became the first government minister to publicly undergo an AIDS test.

Michael Cherry

Oil firms back AIDS project in the Niger delta

BONNY ISLAND, NIGERIA

Oil and gas companies have set up a three-year project to tackle the rapid spread of AIDS around one of the largest facilities in the Niger delta.

King Edward XI of Bonny Island announced the partnership on 1 December, World AIDS Day, saying he hoped it would “rouse the community into action” and lead to free treatment for people living with HIV/AIDS. The island, with a population of about 150,000, houses a vast US\$15-billion liquefied natural-gas plant.

The Bonny Island initiative, named Ibani-se, is initially being

supported by Nigeria Liquefied Natural Gas, Shell and Exxon-Mobil to address the welfare of the local community. The companies spent \$600,000 to start it this year and are expected to put in another \$1.4 million in 2007.

The initiative is led by Donald de Korte, a physician and consultant to drug company Merck, who used to run the African Comprehensive HIV/AIDS Partnerships in Botswana. The plan is to consult local people and establish Ibani-se as a community-based project funded not just by oil and gas companies but by sources such as the Global Fund to Fight AIDS, Tuberculosis and Malaria.

The challenges of operating in the lawless and volatile delta mean little has been done so far to confront the menace of AIDS. High infection rates are fuelled, researchers suspect, by rampant sexual interaction between oil workers, other migrant labour and the many sex workers who surround the oil and gas facilities. Project officials chose Bonny Island because it provides a distinct entity for monitoring and treatment, and is stably governed by the king. It is one of the few places in the delta secure enough for such an initiative to work.

A survey of 800 households

suggests that the AIDS prevalence on the island is around 8%, almost twice the national average. The project plans to counsel and test 30,000 people for HIV over three years, and offer antiretroviral treatment at Bonny General Hospital, the island's dilapidated medical facility. “The first priority is to equip and staff the hospital properly,” says Douglas Pepple, its senior physician.

The project's organizers hope to have 50 patients on antiretroviral drugs by the end of 2007, and 800 (a quarter of those estimated to need them) by the end of 2009.

Colin Macilwain

Purdue attacked over fusion inquiry

Nine months after serious allegations were levelled against high-profile 'bubble fusion' research at Purdue University in West Lafayette, Indiana, the institution is being criticized by researchers within and outside its walls for its apparent failure to respond.

Lefteri Tsoukalas, head of the nuclear-engineering school where the work was carried out, resigned his position in October, expressing disappointment over the slow pace and secrecy of the university's response. Last week, he made his concerns public. In an open letter from his lawyer to Purdue's provost Sally Mason, Tsoukalas called for the university to release an interim report on its progress. "Purdue is a great public university, not a private club," he told *Nature*.

In response, Purdue spokesman Joe Bennett confirmed for the first time that there is an ongoing inquiry at the university. "We are not going to have any comment while this is under way," he said.

The allegations relate to the work of Rusi Taleyarkhan, a professor of nuclear engineering at Purdue. Since 2002, Taleyarkhan has published three major papers claiming to have achieved nuclear fusion by using sound waves to collapse bubbles in deuterated liquids, raising the prospect of a virtually unlimited energy source¹⁻³. Tsoukalas' letter comes days after the announcement that a Texas physicist, working with Taleyarkhan in his lab, has achieved similar results (see 'Bubbling up').

In March, *Nature* reported concerns over the validity and reliability of Taleyarkhan's work⁴, based on statements made by members

of Purdue's nuclear-engineering school, as well as an analysis by physicist Brian Naranjo at the University of California, Los Angeles. Naranjo's study⁵ showed that the spectrum of neutrons produced by Taleyarkhan's experiments fits not fusion, but the radioactive decay of californium, a standard lab material.

Purdue quickly announced that a university panel would review Taleyarkhan's work and make the findings public. In June, it said the review was complete but that its findings and any future steps would remain confidential. Since then, Purdue has released no information (until Bennett's comment this week) about whether or not it is investigating Taleyarkhan. Tsoukalas says he decided to write to Purdue because he had received no adequate response to his concerns since he first expressed them in March. Purdue's policies suggest a timeline of three months for investigating misconduct allegations.

Other researchers in the field are also far from impressed by the apparent lack of progress. Ken Suslick, for example, a professor of chemistry at the University of Illinois at Urbana-Champaign, sent a confidential note in June to Purdue's associate vice-president for research, Peter Dunn, stating that he believed Taleyarkhan's research claims are fraudulent. Suslick detailed a number of reasons for his view, including Naranjo's analysis; the fact that other teams were unable to repeat the work; and an experimental demonstration he attended at which he believed data were being "cherry-picked"

by Taleyarkhan. Suslick has received no reply from Purdue.

"They have to be careful what they say in case the accusations are false, so I'm not horrified they didn't respond," Suslick says. "But it is in keeping with their other foot-dragging. At some point they have to say that they have had an investigation and that they either exonerated him or didn't."

Internally, misconduct allegations have been made that centre on the claim that Taleyarkhan wrote up work with his postdoc Yiban Xu claiming to have achieved bubble fusion, then left his own name off the resulting paper (he is listed in the acknowledgments)⁶. Taleyarkhan cited this study in a later paper³ as independent confirmation of his work.

Taleyarkhan has indicated that he is unable to respond to requests for comment about the various allegations because this might violate university's confidentiality procedures. But Taleyarkhan's co-author, Richard Lahey of Rensselaer Polytechnic Institute in Troy, New York, has defended him. "As far as I have been told, [the data] were taken by the authors of this paper," Lahey says.

Concerns have been raised, however, over the fact that data published in Xu's paper are apparently identical to separate data reported by Taleyarkhan. Xu included a figure showing microphone measurements taken during a fusion experiment at Purdue. It includes data that look identical to those reported by Tale-

"I'm not horrified they didn't respond, but it is in keeping with their other foot-dragging."

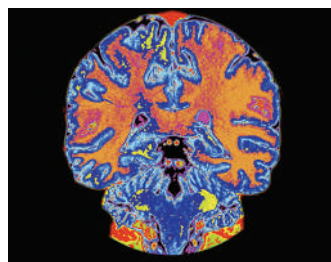
Actions speak louder than images

NEW YORK

Can brain scans of a racist, liar or psychopath accurately tell whether that person will persecute, fib or kill? No, say experts in the ethics of neuroscience, who are increasingly concerned that such images will be used to make dangerous legal or social judgements about people's behaviour. They say it is time for scientists, lawyers and philosophers to speak up about the limitations of such techniques.

"Lawyers want to know 'Can I put somebody on the scanner

and tell if they're racist?'" says Elizabeth Phelps, a psychologist and neuroscientist at New York



Handle with care: can brain scans really identify antisocial people?

University who has studied the brain's response to race. "We as a group of scientists have to be able to say that we can't make that distinction," Phelps spoke at a panel discussion on the emerging field of neuroethics held in New York last week.

Neuroscientists increasingly use technologies such as functional magnetic resonance imaging to see how blood flow in the brain changes when we see pictures, recall memories or make decisions. But these images are prompting

concerns about how they might be over-interpreted or misused (see *Nature* 435, 254-255; 2005).

Outside the lab, neuroimaging is being touted as a way to detect lies (see *Nature* 437, 457; 2005) or to predict what shoppers might buy. There have been suggestions that brain imaging could be used to screen police officers for race bias by showing them faces of particular ethnicities.

But most scientists say that studies of behavioural or physical responses — for example, a person's



Rusi Taleyarkhan's claims to have achieved 'bubble fusion' have been met with disbelief.



**UNANSWERED
QUESTIONS AFTER
RUSSIAN SPY POISONING**
What is known about
polonium-210?
www.nature.com/news

Bubbling up

A physicist at a small Texas college says he has reproduced bubble fusion with the help of Rusi Taleyarkhan.

The work was presented last month at the American Nuclear Society's meeting in Albuquerque, New Mexico, by Edward Forringer of LeTourneau University — a small, evangelical Christian school in Longview. The results were obtained in Taleyarkhan's lab at Purdue University using his equipment, and Forringer believes they tentatively confirm that bubble fusion is occurring.

Forringer went to Taleyarkhan's lab in May after approaching him about his work. Working closely with Taleyarkhan, Forringer says he reproduced Taleyarkhan's earlier results. Forringer adds he did not see any sign that the neutrons he detected might have come from a californium source, as Taleyarkhan's critics have suggested, rather than fusion. But he agrees that reproducing the work in a different lab is what is needed. "We're certainly going to try to do that," he says.

Geoff Brumfield

yarkhan from a prior experiment carried out at Oak Ridge National Laboratory in Tennessee⁷, where he previously worked, as well as those presented by him in a slideshow for DARPA, the Pentagon's research agency, in 2005. The results from Oak Ridge and Purdue were reported to have been produced under different experimental conditions, and the papers have no common authors. Xu has defended the measurements as his own, but declined to say who wrote the paper.

Nature showed the data to three scientists from different groups in the field. All con-

cluded that the measurements must have been taken from the same experiment, because different acoustic cells tend to have characteristic outputs. Lahey disagrees. "If the test sections were of the same design (which they were), the response during cavitation events will be essentially the same," he says.

For Tsoukalas, the lack of word from Purdue is damaging the reputations of all concerned. In the letter to Mason, Tsoukalas's lawyer, Philip Michael, writes: "Significant time has now elapsed since the Purdue Administration made any public statement on the progress of

its investigation, inquiry or examination of the allegations of misconduct which Professor Tsoukalas and other researchers have made... we believe that it is appropriate to ask for an interim report."

Eugenie Samuel Reich

1. Taleyarkhan, R. P. *et al. Science* **295**, 1868-1873 (2002).
2. Taleyarkhan, R. P. *et al. Phys. Rev. E* **69**, 036109 (2004).
3. Taleyarkhan, R. P. *et al. Phys. Rev. Lett.* **96**, 034301 (2006).
4. Reich, E. S. *Nature* doi:10.1038/news060306-1 (2006).
5. Naranjo, B. *Phys. Rev. Lett.* **97**, 149403 (2006).
6. Xu, Y. & Butt, A. *Nucl. Eng. Des.* **235**, 1317-1324 (2005).
7. Taleyarkhan, R. P., Lahey, R. T. & Nigmatulin, R. I. *Multiphase Sci. Technol.* **17**, 191-224 (2005).

reaction to different races in real life — should trump imaging every time. That's because interpreting brain scans, and correlating them to actions, is inaccurate at best. All we can really gain from such studies is a more nuanced understanding of behaviour, says Phelps.

The persuasiveness of brain scans has already drawn them into the court-room. In a landmark case in the US Supreme Court in March 2005, several leading scientific groups, including the American Medical Association, the American Psychiatric Association and the National Mental Health Association, filed briefs to support the premise that teenagers are less

rational than adults.

The data included a brain-imaging study showing that the prefrontal cortex, which governs impulse control and reasoning, develops late in adolescence (see *Nature* 442, 865-867; 2006), and could explain some irrational aspects of teenage behaviour.

Many groups thought this study could help rule against the death penalty. But although the court ruled against the death penalty for those younger than 18, it chose not to cite the brain-imaging study, relying instead on behavioural studies that showed adolescents are more impulsive, more vulnerable to peer pressure and more affected by stress.

Stephen Morse, a professor of law and psychiatry at the University of Pennsylvania, Philadelphia, thinks it was a wise decision. Although the imaging study helped to explain why a particular group might have different behaviour, the behaviour itself is more important than the changes seen in the brain, he says. Citing the study would have given it too much credibility, he adds, and opened the door for further claims that imaging predicts behaviour. "The legal and moral claims being made [from imaging studies involving very few people] are far too extensive."

Morse is a founding member of the Neuroethics Society, set

up earlier this year by a group of lawyers, philosophers and scientists to address issues raised by the use of brain scans and other future applications of neuroscience (see *Nature* 441, 907; 2006). Many neuroscientists are concerned about inappropriate applications of their research, but they rarely come out and say so. Scientists should speak up, but it will also take lawyers and sociologists to lay out the concerns, says Morse. "We need scientists to say what they know and what they don't know," he says. "But the implications are not a science question, they are a moral question, a social question, a legal question." ■

Apoorva Mandavilli

Pfizer to get rights to drugs discovered at Scripps

Drug giant Pfizer will pay US\$100 million over the next five years for first rights to develop drugs from discoveries at the non-profit Scripps Research Institute.

The deal replaces an expiring one between the Scripps facility in La Jolla, California, and Swiss company Novartis. Critics have argued in the past that consumers and government insurance programmes end up paying high prices for drugs generated by taxpayer-funded research. Advocates say the pacts are an efficient way of bringing therapies to market.

The US National Institutes of Health, which funds about three-quarters of Scripps' research at a cost of \$230 million a year, must still approve the new pact. Pfizer would be able to develop drugs based on nearly half of Scripps' discoveries at its California and Florida labs.

GM potato rewrites rules for making paper

The European Union (EU) is considering ending its eight-year effective moratorium on the approval of live genetically modified crops. But on 4 December, an EU panel of experts could not agree on whether to authorize a potato strain that is tailor-made for the paper industry.

The decision will now pass to EU ministers, probably next spring. If approved, the strain, called Amflora, will be the first new commercial transgenic crop to be approved for planting in Europe since 1998. By contrast with conventional potatoes — which produce two forms of starch, amylopectin and amylose — the transgenic version produces only amylopectin, which is more useful for making paper and adhesives. The potatoes would not be grown for eating.



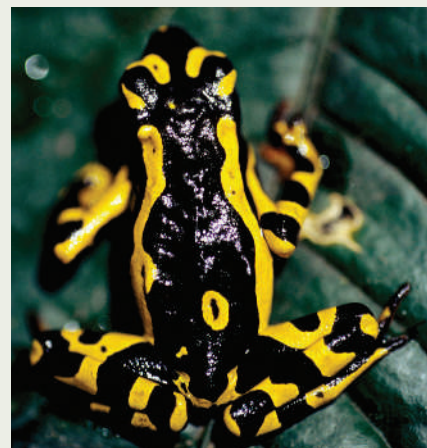
The starch in these potatoes makes them good to eat, but that in a new GM strain will have other uses.

Conservation areas in Brazil set to grow

Brazil has just become a lot more protected. Nine chunks of land totalling the size of Florida have been set aside for conservation.

The government of the state of Pará, working with international conservation groups, announced the new reserves on 4 December. The Gordon and Betty Moore Foundation, based in San Francisco, California, is helping to foot the total cost.

The 16.4 million hectares set aside include the world's largest tropical forest reserve. This is a jungle landscape the size of Denmark, home to endangered species such as the Pebas stubfoot toad (*Atelopus spumarius*; pictured). This region will be open only to research and conservation activities, but other regions will be allocated for sustainable-use projects.



E. BERNARD/CI-BRASIL

Amflora is made by the German chemical company BASF. In a separate development, environmental groups have been protesting against the UK government's decision to approve experimental planting of a food potato engineered by BASF to resist blight. That crop is already in field trials in Germany, Sweden and the Netherlands.

Europe feels a record autumn heat

Been feeling warm recently? Europe has had the hottest autumn in the past 500 years, a Swiss climate historian argues.

Preliminary analysis suggests that mean temperatures over continental Europe have been nearly 2 °C higher than the long-term average for the period from September to November. This autumn was even 1 °C warmer than the record-warm autumns of 1772, 1938 and 2000, says Elena Xoplaki of the University of Bern, who led the unpublished study.

"Exceptionally warm autumns in one region or another wouldn't be so telling," says Jürg Luterbacher, a climatologist at the University of Bern. "But the signal is consistent over the whole European land mass, from Iceland to Greece."

Autumn climate trends have been generally less well investigated than summer, winter and spring trends.

Agricultural priorities adapt to climate change

Faced with the prospect of climate change devastating crops worldwide, agricultural researchers have tweaked their research agenda.

At its annual general meeting in Washington DC this week, the Consultative Group on International Agricultural Research — an international alliance of

food-crop centres — released several dire predictions for the future. For example, heat and drought are expected to shrink the wheat-growing area of India by half by 2050. Elsewhere, though, climate change would open new areas for crops: in North America, for instance, wheat could be grown at latitudes nearly up to the Arctic Circle.

But most of the world's poor live in the tropics, where food-growing capacity will be hardest hit. So the centres have laid out new research priorities, which include breeding improved crops to better tolerate salt, drought, flooding and rising temperatures, and helping farmers manage their resources more effectively.

Conflicts of interest found on hospital review boards

Even the people who serve on institutional review boards for hospitals often have financial ties to industry, a study shows.

Of 893 board members surveyed, just over a third reported having at least one such relationship. And 15% said that, within the past year, an issue had come before their board that involved either a company in which they had a financial relationship or a competitor of such a company. Of those, 23% said they never disclosed such potential conflicts of interest to a board official.

The survey, led by Eric Campbell of Massachusetts General Hospital in Boston, appeared last week in *The New England Journal of Medicine* (E. G. Campbell *et al.* *N. Engl. J. Med.* 355, 2321–2329; 2006).

In a related study published in the same issue, a second team found that more than 90% of patients in US cancer trials said they did not worry about any possible financial relationship between industry and the researchers conducting the trials (L. A. Hampson *et al.* *N. Engl. J. Med.* 355, 2330–2337; 2006).



BIOFUELLING THE FUTURE

The idea of using living plants as a way to capture the all-but-unlimited energy of the Sun has a powerful romantic attraction. Unfortunately, plants have a basic problem in this respect. Compared with, say, an array of solar cells, they are strikingly poor transducers of the Sun's energy; even an intensively managed plantation struggles to store away more than a watt or two per square metre on average.

But plants have advantages that, in some circumstances, outweigh their low efficiency. For a start, unlike solar cells, plants are very cheap to make; indeed, with a moderate supply of water and nutrients they will make themselves. They use up carbon dioxide in the process, which is a definite environmental plus — and they turn that carbon, along with the Sun's energy, into stable organic compounds.

This means that the Sun's energy is made available at a later date when the Sun isn't shining. And it means that it can be processed, at some cost and effort, into hydrocarbon fuel of the sort that today's cars, trucks and planes can handle with relatively little modification. To governments worried about the stability of their fuel imports, or indeed the long-term future of global oil production, that could matter quite a lot.

So although their inefficiency means that plants will never be the total answer to our global energy problems, they have substantial potential as a source for carbon-neutral fuel for the ever-thirsty transportation

sector, as long as oil prices stay reasonably high.

Our Features this week look at this potential in three different areas. The first (see page 670) addresses the world's most substantial biofuel success — the Brazilian sugar-cane ethanol industry — and assesses its impact and potential. The second (see page 673) looks at the possibilities of making ethanol, or other alcohols, out of sources of cellulose from farm waste to poplar plantations — possibilities currently entrancing US entrepreneurs.

The third feature (see page 677) looks at a different approach to alternative fuels — the thermochemical route. This technology can be used to make fuel from biomass — but it can also, more easily, be used to make liquid hydrocarbons from solid coal, and this is where most research in the area is focused. For countries that have a lot more coal than oil, this coal-to-diesel technology can look attractive purely on an energy-independence basis. But without careful and costly sequestration of the carbon dioxide produced, it would be a significant additional source of greenhouse gases.

No fuel technology is perfect. But, as we argue in our Editorial (see page 654), a greenhouse-gas crisis and worries over oil supply mean that diversifying the range of options makes good sense — as does the development of new routes from research to large-scale deployment. ■

J. BLAIR/CORBIS

Drink the best and drive the rest

Brazil's sugar-cane ethanol industry is the world's best and able to get better, says **Emma Marris**.

In Brazil, alcohol made from sugar cane is mixed with lime juice and a little of the cane sugar itself to make caipirinhas — and it's a fine way to get the weekend off to a flying start. But come Monday morning, sober Brazilian commuters are still using cane liquor — in their fuel tanks. Most Brazilian petrol is gasohol, which by government mandate is currently 23% ethanol. Next to the gasohol pumps at the petrol stations are pumps that offer pure ethanol.

Ethanol from sugar cane has been powering cars in Brazil on and off since the 1930s, and with government backing since the OPEC price rises in the 1970s. It makes fairly obvious sense. Brazil's tropical sun makes it a great place for growing sugar cane: it is the largest cane producer in the world, producing more than twice as much as the number two, India. Just crush out the sucrose solution, ferment it into alcohol with the help of yeast and distill it to the desired concentration; burn the 'bagasse' — the fibrous pulp left over when the sugar is squeezed from the cane — to power the process along. Put the alcohol into your gas tank and you are effectively driving it on sunlight.

In the past few years, Brazil's bioethanol industry has sprouted new wings thanks to higher petrol prices and 'flex-fuel' cars, which can sense different mixtures of petrol and ethanol and adjust their workings accordingly. Introduced to the mass market in 2003, these cars changed everything; flex-fuel vehicles now account for well over half of Brazil's new cars. Their attraction is that they allow the owner to trade off continually between the advantages of neat ethanol — which gives 20% to 30% fewer kilometres to the litre — and gasohol depending on the current prices and the local tax rates. (Tax alone means that if you were to take a road trip across Brazil, you'd switch back and forth a couple of times as you crossed from state to state.)

Sweet solution?

Flex-fuel cars have created an expanded home market for Brazil's ethanol, and production is up. The country produced 282,000 barrels (45 million litres) a day in 2005, up from 192,000 in 2001. The ministry of agriculture is projecting 442,000 barrels a day by 2010. Oil is still very much ethanol's big brother, and in the first part of 2006 the country was producing an average of more than 2 million barrels a day while consuming a bit more. But 40% of the fuel powering Brazil's cars is home-grown ethanol. The sugar cane industry supports more than a mil-



PAULO WHITAKER/REUTERS

Cleaning up: Brazil's use of sugar cane, here being washed for refining, significantly reduces CO₂ emissions.

lion jobs, in a country with an unemployment rate of 10%. Some bagasse-powered mills sell surplus power back to the grid. And although the carbon put into the atmosphere by petrol-powered cars comes out of fossil reserves, the carbon from ethanol is carbon that was in the atmosphere just a couple of years ago, before the sugar cane got hold of it and worked its photosynthetic miracle. There are thus, in principle, no net emissions.

This all sounds a bit like a free lunch, or at least a free ride, and everyone knows they don't exist. Growing sugar cane, harvesting it, fermenting and distilling it and then distributing it is a complex business. It uses inputs — fuel for harvesters, nitrogen fertilizers for the cane — that themselves require energy from elsewhere. And it has potentially damaging side-effects, such as soil erosion and the emission of nitrous oxide, a greenhouse gas, from farmland. Taking all the complexities into account, is ethanol still as good a deal as it seems? And if it is, how much of this good thing can Brazil — and the rest of the world — get its hands on?

In 2004, Isaias de Carvalho Macedo at the University of Campinas did a study for the state of São Paulo that considered energy inputs such as fertilizer manufacture and agricultural

machinery in the sugar-cane industry¹. He and his colleagues estimate that the whole shebang costs about 250,000 kilojoules per tonne of cane. That tonne of cane in turn yields about 2 million kilojoules in ethanol and surplus electricity made by burning bagasse. That's an eight-fold return.

This is a lot better than ethanol-makers in the United States manage, and the reason is clear to anyone who's ever strolled through a cane field chewing on a bit of the stuff; cane is a far more prolific plant than corn (maize), from which the United States makes almost all its ethanol, and it puts a great deal of its — or rather the Sun's — energy into making sugar.

What's more, sugar cane needs less by way of inputs, and in the parts of Brazil where most of it is grown at the moment it needs no irrigation. It needs only to be ploughed up and replanted every five years; between times it can be cropped repeatedly and will simply grow back, although the yields drop a bit with each harvest. For all these reasons, sugar-cane ethanol is also currently the cheapest ethanol to produce in the world. A litre costs about 25 cents to make. The commodity price for anhydrous ethanol (the kind mixed into gasohol) is about 27 cents.

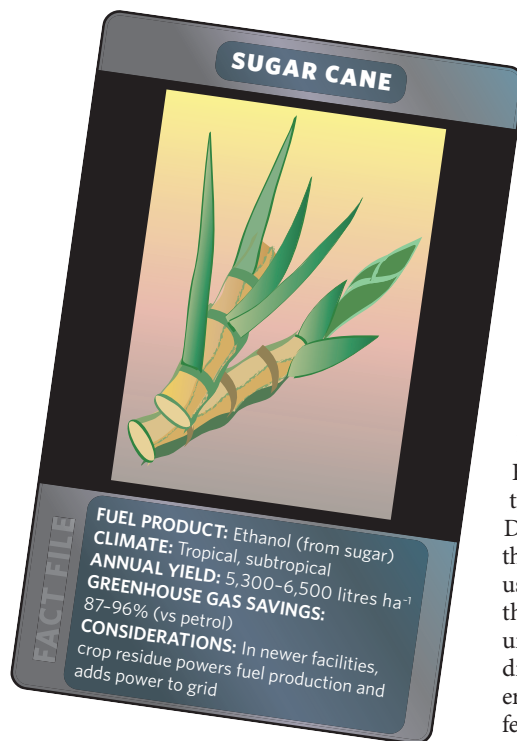
Because of this a lot of money is pouring

into the centre-south region of Brazil, where the sugar cane grows best. But that does not necessarily mean that investors believe Brazilian cane ethanol is sure to become the planet's biofuel of choice. A range of other emerging liquid-fuel technologies, including biodiesel, butanol, coal to liquid, and the buzz leader at the moment, cellulosic ethanol (see page 673), are attracting investment around the world, too. And the investors are typically backing a range of possibilities while they wait to see how the market pans out.

Arnaldo Vieira de Carvalho, an energy specialist at the Inter-American Development Bank in Washington DC, sees current investment in bioethanol as, among other things, a way of building up a better stake today for whatever the most impressive biofuel technology turns out to be tomorrow. "If you now put your money in distilleries, in five years you have made your money," says de Carvalho, "and then you put your investment in the technology that is coming."

Fuelling change

As the rest of the world gets interested in ethanol, both as a fuel in its own right and as a fuel additive (to replace methyl tertiary-butyl ether, or MTBE, an anti-knocking additive that is on its way out for environmental reasons), Brazil is ramping up exports (see 'Key companies'). Brazilian manufacturers are actively promoting ethanol distribution systems in other countries. Petrobras, until the 1990s the state oil monopoly and



still the biggest player in Brazilian oil, is planning to build an ethanol pipeline from the centre to the south, out to the ports. It will be the first of its kind. Pipelines for ethanol have long been considered problematic, as it tends to absorb moisture and impurities as it flows. Construction will start in January, according to Plinio Mario Nastari, president of Datagro in São Paulo.

Not all foreign markets are easily tapped. Increasing Brazilian ethanol imports to the United States led the United States to slap on a tariff of 14 cents per litre (54 cents per gallon), to protect its highly subsidized corn ethanol makers. One effect of this has been to put Jeb Bush, Florida's governor, at odds with his brother President George W. Bush; the governor would rather buy cheap ethanol from Latin America than expensive stuff from the Midwest.

Environmentalists are watching all this expansion carefully. The Macedo analysis suggests that a tonne of cane used as ethanol fuel represents net avoided emissions equivalent to 220.5 kilograms of carbon dioxide when compared with petroleum with the same energy content. The team extrapolates that ethanol use in Brazil reduces greenhouse-gas emissions by the equivalent of 25.8 million tonnes of carbon dioxide equivalent a year. For comparison, Brazil's total carbon dioxide emissions from fossil fuels is 92 million tonnes a year, according to the US Department of Energy. The improvement is thus substantial, if not out-of-this-world.

But there are environmental worries that go along with the ethanol industry too: as well as fertilizer and fuel use, there are also pesticides and pollutants

such as liquid waste and smoke from burning fields to take into account. Last year, a paper by a group at Washington State University in Richland made headlines by claiming that in this broadest perspective, Brazil's ethanol was bad for the environment².

This result, though, is disputed — and the industry seems to be getting greener as it goes hunting for efficiency gains. As Christopher Flavin, an analyst at the World Watch Institute, a green pressure group in Washington DC, points out, expansion will generally mean that a higher percentage of the industry will be using newer, cleaner technology (although by the same token it will also be offering fewer unskilled distillery jobs). The leftovers from distillation, once commonly thrown into rivers, are now often spread on fields as potassium fertilizer. Field burning, which by scorching the cane makes it easier to harvest with machetes, is decreasing both as a result of legislation and the increased use of mechanized harvesters. "Thirty-five per cent of cane harvesting is already mechanized," says Nastari, "and it will increase," although because the machines work only on very flat land some burning is likely to continue. Research on planting methods that involve less or no tilling should lead to reduced erosion, too.

Perhaps the biggest environmental worry is that expanding ethanol production will lay waste to natural forests in its path, reducing biodiversity and releasing stored up carbon.

The first thing to remember on this





Interior designs: some of Brazil's Cerrado region could be cleared for sugar-cane plantations.

front, say the Brazilians, is that Brazil is big. It is big enough to expand its sugar-cane fields massively without either displacing necessary food crops or getting anywhere near the rainforest that the rest of the world seems to have decided is international property. São Paulo state itself is as big as the United Kingdom. Although Brazil is already by far the largest producer of sugar cane in the world, sugar cane is only its fourth biggest commodity, in terms of revenue, with cattle, chicken and soya all bringing in more money. The limiting factor in expansion is capital rather than land.

Growth area

According to Nastari, cane fields account for just 5.7 million hectares in a country of 850 million hectares. There are already 100 million hectares of old agricultural land or pasture land in the centre-south available for the industry to expand into. Eduardo Pereira de Carvalho, president of Unica, the union of cane-growers in São Paulo, welcomes such expansion with open arms: "As for conflict between food and energy, the fantastic increase in productivity has made all these Malthusian arguments completely nonsense, and we have hundreds of millions of hectares of idle land."

Expansion into at least some of those millions of hectares would probably be more or less carbon-neutral, says Robert Boddey, a soil chemist at Embrapa, the Brazilian agricultural research unit. "For degraded pastures, which are slowly losing carbon, it is not such a bad change. And almost 70% of the Cerrado [Latin America's savanna, of which Brazil has some 200 million hectares] has already been cleared." Because it needs a dry season, sugar cane would not be a good crop to move into cleared rainforest

areas, even if that was what anyone wanted. In this respect, cane is more environmentally friendly than palm oil, the most energy-intensive source of biodiesel. Palm-oil plantations for biofuel are having serious effects on the primary forests of Indonesia, but are not as yet big business in Brazil.

Overall, the environmental problems with cane production, says Flavin, are "dwarfed by the land use issues posed by soyabeans and cattle". But they may get worse — or exacerbate the problems elsewhere in agribusiness. Jason Clay of environmental group the WWF points out that "the push of sugar cane is going to displace other crops, and some of

Key companies

Cosan, based in São Paulo, is the largest sugar-cane producer in the world, and the second-largest producer of ethanol. It has been gobbling up smaller ethanol mills right, left and centre, which has given it a very respectable growth curve. In the 2005–06 crop season, it reported revenues of US\$1.1 billion and produced about 8.7 million barrels of ethanol. Cosan went public in November of 2005 at about \$19 a share, and prices were close to \$60 this summer.

Copersucar, a cooperative of sugar and ethanol mills, was founded in 1959. In the 2005–06 crop season, its revenues were \$2.1 billion, and it produced 23 million barrels of alcohol.

them can be grown in the Amazon — cotton, soya, livestock". Some of that soya could be for biodiesel use.

Alex Farrell, head of the Energy and Resources Group at the University of California, Berkeley, adds that for all the enthusiasm there is still a dearth of evidence on the long-term sustainability or otherwise of cane farming. But, he adds, that is pretty much par for the agricultural course: "All around the world, every government has agriculture — they never ask 'is this agriculture sustainable?' Not for sugarcane, not for rutabagas."

To replace a tenth of today's global petrol production, Brazil's ethanol production would have to grow by a factor of 40 or so. Few see that as likely in itself. Even Unica's de Carvalho, who is undeniably bullish, sees only a doubling by 2014, though he does not see that as the end of the story. Enthusiasts for new cane varieties talk of doubling the yield per hectare, but not necessarily going much further.

But a doubling or two is not to be sniffed at, and there is increasing interest in spreading the techniques developed in this most cane-friendly of countries to others in Latin America and elsewhere, with the details changed to fit local conditions. Already, ethanol is becoming a large enough business for the price of sugar on world markets to respond to changes in the oil markets — so the price of your caipirinha is now, in a very small way, susceptible to the manoeuvres of OPEC.

Emma Marris is a reporter for Nature based in Washington DC.

1. Macedo, I. et al. Assessment of Greenhouse Gas Emissions in the Production and Use of Fuel Ethanol in Brazil (prepared for the Government of the State of São Paulo) (March 2004).
2. Dias de Oliveira, M. E., Vaughan, B. E. & Rykiel, E. J. *Bioscience* **55**, 593–602 (2005).

See Editorial, page 654.



A field in ferment

To move US biofuels beyond subsidized corn will be a challenge, reports **Katharine Sanderson**.

Critics of the US ethanol industry have long derided it as an environmentally questionable subsidy to Mid-western farmers that simply serves a transparently political purpose. Voters in Iowa, the buckle in the US corn belt, get first say in the process of choosing presidential candidates. All such candidates are in favour of turning corn (maize), which the state produces in abundance, into ethanol. This pre-presidential support is good for the Iowan economy, but not necessarily that great for the environment.

Studies that compare the energy that goes into making ethanol — expended during the harvesting, fertilizing and transporting of the corn to refineries, and then refining it — with the energy that is released when it is burned routinely show that the net gain is at best small. The American Coalition for Ethanol says that ethanol contains twice the amount of energy that is used to make it; critics see no net gain whatsoever.

This criticism has had little effect, and since 1980, US ethanol production has risen from an average of 6,500 barrels (1 million litres) a day to 260,000 barrels a day. Federal mandates call for a further doubling by 2012. But it is increasingly clear to many in the industry that the criticisms of corn-based ethanol have merit, and in 2006, the need for an alternative was given the highest profile it could get when President George W. Bush brought it up in his state of the union address. In order to improve US energy security, he said, his government intended to make cellulosic ethanol (ethanol made from the rougher and woodier parts of plants) a competitive biofuel within six years.

Corn stores

The advantage of an ear of corn as a source of ethanol (or for that matter as a bit of food) is that it is mainly starch, which is made up of sugars linked in a regular way with bonds that can be broken easily. Breaking the bonds between sugars and using yeast in the fermentation to produce ethanol is a straightforward task for the biorefineries. The disadvantage is that corn is a crop that needs a lot of inputs — fertilizers, water and pesticides — and that doesn't put as much of the sugar it creates through photosynthesis into its ears as one might wish. A lot of the sugar is instead turned into stalks and 'stover' — structural material rich in cellulose and considerably more difficult to break down.



Plant processing: logen's enzyme fermentor, a later stage of turning cellulosic feedstock into ethanol.

Plants that store up a significant amount of energy in easily usable forms such as starch or sugar are exceptions, encouraged in their oddities by millennia of selective breeding — and of them all, only sugar cane grown in the tropics puts enough energy into its easily purified products to make bioethanol obviously attractive (see page 670). Most plants put the bulk of the energy they store up from the sun into cellulose and a related polymer, hemicellulose, and woody plants add another substance, lignin, to the mix. Cellulose makes up the plant's cell walls and, like starch, it is a polymer of sugars containing six carbon atoms linked one to the next. Hemicellulose, on the other hand, is based on a five-carbon sugar, xylose, although it contains many other sugars as well; its various components are thrown together in messy looking chains with many branches. Lignins are huge crosslinked jumbles of organic molecules

which reinforce cellulose and hemicellulose to turn them into wood.

The energy that the plants put in to making the bonds in these various substances could, in principle, be extracted by fuel makers. And these molecules — particularly cellulose, which is both the most abundant and the easiest to dismantle — are much more plentiful than starches and sugars. But they are also much harder for microbes to break down; if they weren't, there'd be no trees, just pools of green goo. As yet, there are no cellulosic ethanol refineries operating at full commercial capacity, and assessments of the technology's readiness for market vary a great deal, as do opinions on how to get there from here. Government incentives and tax breaks might be one solution, but big energy companies also have a role to play, as do the smaller companies that have already worked on developing

the technology, but have not yet found the best ways of spreading and licensing it.

The most expensive part of making ethanol from cellulose is pretreating the biomass to make it accessible to the enzymes that will then cut the sugars from the polymers so that they can be fermented. Typical pretreatments reduce the feedstock's volume chemically using acids, peroxides and ammonia, often along with some form of mechanical pressing or shredding. Unfortunately, this is not a step that can be skipped to cut costs, says Charles Wyman of the University of California, Riverside, because high sugar yields are essential, and untreated biomass gives very low yields. "The only step more expensive than pretreatment is no pretreatment," he says. Instead, the hunt is on for pre-treatment technologies that involve fewer chemicals, require less energy and don't degrade the sugars that are set free in the process.

After the pre-treatment stage comes the snipping out of the sugars, which is the point at which biotechnologists think they can greatly improve on the current process. Abengoa Bioenergy of St Louis, Missouri, a subsidiary of the Spanish engineering group Abengoa, recently invested \$10 million in Dyadic Interna-

tional, a biotechnology company that is concentrating on enzymes for degrading cellulose.

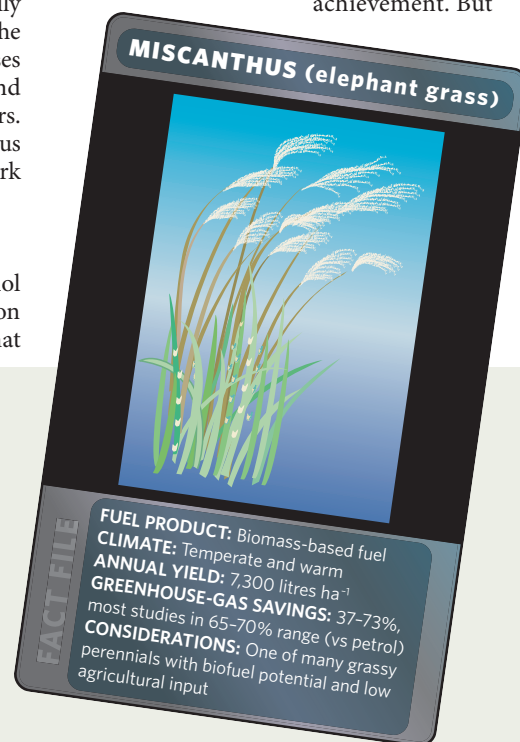
Based in Jupiter, Florida, Dyadic didn't start out as an energy company — in the 1970s it was a leading supplier of pumice for stone-washing jeans. But the enzymatic expertise it developed for distressing denim was then turned to a number of other ends. One of those was breaking down wood, a job that in nature largely falls to fungi. The company's research has centred on a filamentous mess of a fungus discovered by accident in a Russian forest that now, after ten years of processing and genetic engineering, makes up Dyadic's patented C1 fungal cell system. The fungus has been fully sequenced and encouraged to overexpress the genes that then make cellulases and xylanases — the proteins that break up cellulose and hemicellulose to produce fermentable sugars. "We have the world's most prolific filamentous fungus," boasts Dyadic's chief executive Mark Emalfarb.

Cellulose solutions

Emalfarb believes that the cellulosic ethanol market could eventually be worth \$20 billion a year in the United States, and suggests that

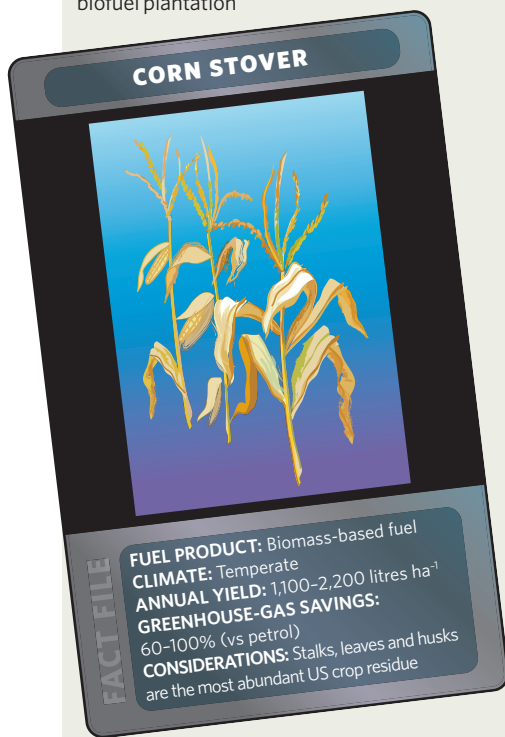
there is enough raw material available in the United States to produce 2.4 billion barrels of cellulosic ethanol a year. This is a bit more than half of what some estimates claim is needed to completely replace petrol as a fuel — the United States gets through some 3.3 billion barrels a year, but the energy content of ethanol is lower than that of petroleum.

The current leader in the cellulosic ethanol market, Iogen, also uses fungal enzymes. The company makes small commercial quantities of ethanol from straw at its pioneering cellulosic ethanol facility in Ottawa, Canada. As the first of its kind, this is an undoubted achievement. But



Prairie dreams

Endless hectares of grass, some tall, some short. Goldenrod, purple clovers and thistly, milky weeds, all growing together in all-but-virgin prairie. It doesn't sound much like a biofuel plantation



— but some researchers say such mixtures could produce as much fuel as any alternative, and with a very low ecological impact.

"We have to find a way for people to see biomass as a renewable energy crop that won't compete with food or affect commodity prices," says David Tilman, an ecologist at the University of Minnesota in St Paul, and he thinks that mixed planting might be the way. Experimental plots containing mixtures of prairie plants are, perhaps unsurprisingly, more resistant to drought and pest than farmed monocultures¹. Rather more surprisingly, they yield 2.7 times as much biomass on average², even outperforming the highly touted energy crop switchgrass. And in any cellulose-based system, biomass is king: the more there is of it — at least in theory — the more fuel can be extracted from it.

Tilman's prairies are one possible source of biomass, their appeal enhanced, he says, by the fact that the plants can thrive on land considered marginal for agricultural use. But they are not the only source under investigation. From improved strains of trees to plants genetically engineered for easy processing, there is a wide range of options, most of them with advocates who tout their ideas as enthusiastically as Tilman.

Richard Flavell, chief scientific officer of Ceres, a biotechnology company based in Thousand Oaks, California, gives full credit to Tilman's academic work — but says that how his ideas will

play out in the context of biofuels remains to be seen. One issue is that a polyculture approach means finding the right mix of plants for lots of different conditions. Another is that, especially with technologies not yet fully developed, biofuel producers will prefer a uniform feedstock.

The enzymes that convert biomass into fuel are picky and easily inhibited by substances in various plants. That argues for dedicated plantations of predictable, well characterized plants — poplars, perhaps, in wetter climates, switchgrass or miscanthus, also known as elephant grass, in drier, warmer places. These perennial crops represent a significant improvement over corn as a feedstock: compared with corn, switchgrass cultivation requires less fertilizer and water, and results in one-eighth the nitrogen runoff and one hundredth the soil erosion, according to the

even when it reaches its full capacity, which it is taking quite some time to do, it will be capable of producing only 2.5 million litres (16,000 barrels) a year, which is not a great deal.

Iogen chief executive Brian Foody is not worried. The critical steps for getting the right enzymes, the right pretreatment systems and the right yeast systems, have all been done, he says. "We just need to go through the nuts and bolts of the process." This means making sure that the demonstration plant works well enough to be replicated elsewhere — the company is looking to build new facilities in Idaho, Saskatchewan and Germany.

Iogen recently secured a \$30-million investment from the bankers Goldman Sachs, bringing the total invested in it since the 1970s up to \$130 million. But not all potential investors are convinced. "I don't really understand what Iogen is doing," says Matt Drinkwater, market analyst at New Energy Finance in London, UK. And his concerns are not unique to Iogen — many of the companies in the sector, he says, hold details of their processes so close to their chests that they are hard to evaluate, whether they be relatively small outfits such as Iogen or giants such as DuPont, which is also

developing cellulosic ethanol technologies. Robert Wilder, who manages the Wilderhill clean energy index — the first such index to be accepted on Wall Street — agrees, but acknowledges the constraints that the chief executives of small cellulosic ethanol companies work under in terms of not tipping their hands to larger competitors.

Smells like green spirit

Perhaps because of these uncertainties over the technology's readiness, most of the money that has been invested recently in ethanol production both within the United States and beyond has been in the more traditional technologies. The sizable investments being made by agribusiness giant Archer Daniels Midland — the biggest ethanol producer in the United States and, perhaps tellingly, a company run by a chief executive who was recruited from the oil industry — seem mostly to be in traditional corn ethanol. The same applies to high-flying UK entrepreneur Richard Branson's recent investments in Ethanol Grain Processors of Tennessee and a new grain-based Californian ethanol venture, Cilion.

But there is

some evidence that enthusiasm for investing in corn ethanol may be waning. Various ethanol companies that were riding high earlier in the year saw their stock slump after the summer when oil prices came down from their \$78 a barrel peak.

This might mean the market is aware that, although subsidies may be able to keep it profitable for the time being, there is no way that corn ethanol can make a marked difference to long-term energy use in the United States. To make enough ethanol to start seriously displacing oil imports requires a process that can use cellulosic materials such as switchgrass, a tall prairie grass, or miscanthus, a grass imported from Asia, which provide far more tonnes of biomass per hectare than corn kernels ever can, and can be grown on land not suitable for conventional agriculture. Other sources could be farm waste or trees or newly engineered plants of some sort (see 'Prairie dreams'). This leads to something of an investing impasse: the companies in the business at the moment make money; the ones that might take it to the next

stage do not, in large part because no one has made the heavy capital invest-

R. HODGKINS

US Department of Energy. And with current cellulosic technologies, switchgrass could yield roughly 4,000–6,000 litres per hectare, rivalling or beating the mature corn technology.

And that is before Flavell and his competitors get their hands on the stuff. Ceres wants both to enhance the biomass produced and to reduce the inputs needed, producing a crop that flourishes on marginal lands. The company's approach is to identify favourable genes in *Arabidopsis*, the lab rat of the plant world. So far, researchers at the company have found genes that boost biomass, increase nitrogen-use efficiency and increase resistance to the stress of drought, cold or salt.

Ceres has a \$137-million licensing agreement with Monsanto to characterize such genes for new varieties of traditional row crops such as corn and soya bean. It is also using the genes in molecular-marker assisted breeding programmes for switchgrass and other crops in collaboration with the Samuel Roberts Noble Foundation based in Ardmore, Oklahoma.

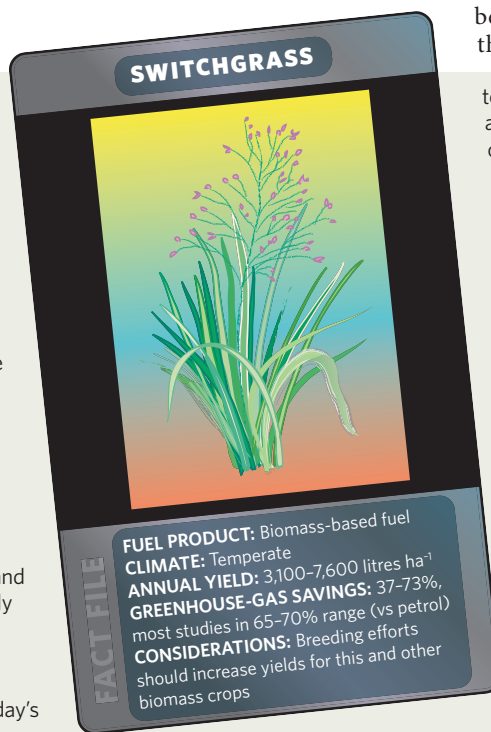
Flavell points out that most biofuel crops have not been selected for high biomass, so breeders have a lot of genetic material to work with. The Noble foundation has already increased the yield of switchgrass by 25%, using conventional breeding.

But the genetic engineers are also laying plans, and looking beyond brute biomass to what could be seen as functionality — developing crops that actually help with the biofuel conversion process. For example, Ceres

and other companies, including Edenspace Systems in Dulles, Virginia, are engineering crops to produce enzymes that would break down their own cellulose when triggered.

In the long run, there is no need to see polycultures and monocultures as mutually exclusive — both could have a role, and both could probably have their yields improved. If mixed feedstocks are not well adapted to today's finicky enzymes, they could still be used by 'thermochemical' systems such as those using the Fischer-Tropsch reaction (see page 677). And in time, the enzymes may catch up.

As Thomas Faust, biofuels research manager at the US National Renewable Energy Laboratory in Golden, Colorado, points out, the fundamental trade-off is between processing and environmental impact. "If you were going



to genetically engineer a plant with desirable characteristics, a monoculture makes a lot of sense," he says. "From a conversion perspective, your ideal feedstock would be very homogeneous, a single grass, with constant moisture, available all year round. But to have a truly ecological, good-for-the-environment option, you would want mixtures." And according to Tilman, recent research on the biomass benefits of biodiversity³ means that such mixtures could be found to suit many parts of the world: "The mechanisms behind

the effects we are observing should apply to all other plant communities."

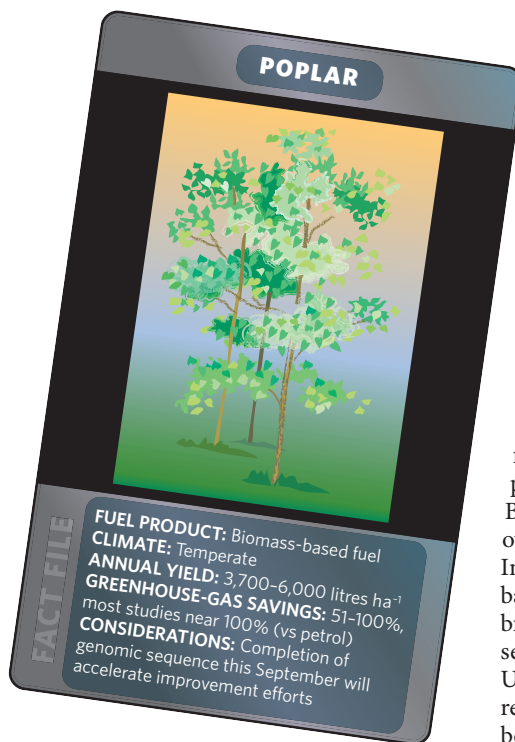
Charlotte Schubert

1. Tilman, D., Reich, P. B. & Knops, J. M. H. *Nature* **441**, 629–632 (2006).
2. Tilman, D., Reich, P. B., Knops, J., Wedin, D., Mielke, T. & Lehman, C. *Science* **294**, 843–845 (2001).
3. Van Ruijven, J. & Berendse, F. *Proc. Natl. Acad. Sci. USA* **102**, 695–700 (2005).

ments needed for plants that make use of the technologies that have already been piloted.

One way round this is to invest across the board. This is the strategy pursued by Vinod Khosla, the Silicon Valley venture capitalist who is one of the founders of Cilion. Khosla is also involved in cellulosic technologies through two companies based in Cambridge, Massachusetts: Celunol, which has just started to operate its own pilot plant, and Mascoma, which concentrates on process engineering and which last month raised \$30 million in second-round venture funding. Farther afield in the biofuels world, Khosla is also a major investor in Kergy, a company that turns biomass into fuel in a completely different 'thermochemical' way, using just heat and catalysts. For some observers, such as Dan Schrag, a geochemist at Harvard University, these approaches are more attractive than fermentation, not least because they need no witches' brews made from fiddly feedstock-specific enzyme. "When the dust clears, cellulosic ethanol is unlikely to be where we end up," he predicts (see page 677).

To Drinkwater, investors such as Khosla, with their broad-based approach to the problem, are exactly what the industry needs to drive the market forwards and get it over the final bump it needs to clear before commercial success. Unfortunately, there are few such people. In their absence, many in the industry, not without self-interest, see the responsibility resting with governments to provide attractive tax incentives. "All forms of energy should



face market prices that reflect the cost to society that they impose," says Foody. And to set those market prices, the right tax incentives and government mandates need to be in place.

But government incentives won't make the scientists any smarter, and observers outside the pioneering companies believe there is still basic work to be done before those companies, or their eventual competitors, make the process economically viable. Thus they welcome increasing levels of basic research from the

government, such as the US Department of Energy's pledge of \$250 million to set up two bioenergy research centres that are largely focused on cellulosic ethanol. The European Union has set aside €100 million (US\$132 million) for cellulosic ethanol in its seventh Framework Programme on research.

R. HODGKINS

Ethanol alternative

Companies large enough to afford it are also following the basic research route rather than placing early bets on particular technologies. BP has announced it will invest \$500 million over ten years to fund an Energy Biosciences Institute, which will be a dedicated facility based at a university. The University of Cambridge, Imperial College London, Massachusetts Institute of Technology, Stanford, the University of California, Berkeley, and Lawrence Berkeley National Laboratory have all been mentioned as possible hosts — the final decision is expected in December.

One intriguing possibility for such research to pursue is replacing ethanol with another form of alcohol. The fact that ethanol is easy to ferment can blind people to the fact that it has almost as many inherent problems as a fuel as corn has as a feedstock. Its tendency to pick up water wherever it goes makes it hard to transport, particularly in pipelines. It's corrosive. It's more volatile than one might wish. And its energy density is low compared with regular petrol.

For these reasons, BP and DuPont are working with British Sugar to adapt their ethanol fermentation facility in East Anglia to produce butanol — an alcohol with four carbons in it, as opposed to ethanol's two. This requires training microbes in new tricks, but it is not as hard a problem as breaking down woody plant material. The East Anglia plant will use locally grown sugar beet as the feedstock, but in the long term the aim would be to use a cellulosic feedstock. "We accept that taking stuff out of the food chain is not the right way to go," says Robert Wine, a BP spokesman.

Drinkwater thinks that an industry demand for butanol as an end product could actually increase interest in cellulosic approaches. "Most refiners would be much happier to use butanol than ethanol," he says. If oil companies become confident in biofuel technologies, investors would in turn be more confident of the biofuels industry as a whole, giving the industry that elusive final shove that it seems to need.

Katharine Sanderson writes on the physical sciences for *Nature*.



Pastures new: miscanthus, or elephant grass, is an alternative energy crop grown to produce ethanol.

See Editorial, page 654.

T. GRAHAM/ALAMY

Making it up as you go along

Chemists can make liquid fuel from biomass — or from coal. **Heidi Ledford** weighs up the pros and cons.

Brian Schweitzer, a rancher and the Democratic governor of Montana, has ‘folksy’ down to a fine art. In a bolo tie, jeans and cowboy boots, Schweitzer shifts seamlessly from jokes about how his border collie, Jag, drinks out of the toilet to analyses of energy policy in which every complex problem gets its down-to-earth soundbite, from pumping carbon dioxide into the ground (“It came from those rocks, we’re just sending it home”) to public perceptions of coal as a dirty source of energy (“To a lot of people, coal is a four-letter word”).

Put these *aperçus* all together and you have a pitch for turning coal into oil, an idea Schweitzer and his constituents have 240 billion reasons to take seriously. With 8% of the world’s reserves, Montana has enough coal to make 240 billion barrels of diesel fuel, which is in the same ball park as all of the proven reserves claimed by Saudi Arabia.

Schweitzer is far from alone; many politicians and businessmen are now eyeing a chemical process called the Fischer–Tropsch synthesis as a way of converting solid hydrocarbons or natural gas into liquid fuel. The great advantage is energy security: Fischer–Tropsch technology allows domestic coal to replace foreign oil, which is pretty attractive if you’re sitting in Washington DC or Beijing, let alone Billings, Butte or Bozeman. Another potential advantage is environmental: the fuel produced by Fischer–Tropsch methods can be made to burn more cleanly than diesel. It could thus ease the adaption of cars with efficient diesel engines in countries, such as the United States, that have so far been resistant to such technology.

The obvious drawback, though, is also environmental. The process of converting coal into liquid and using it for transportation releases nearly twice as much carbon dioxide as burning diesel made from crude oil does. In a world conscious of climate change, that excess carbon is a problem. “If you make liquids from coal and don’t capture carbon dioxide in the process, you’re effectively doubling emissions,” says Eric Larson, a research engineer at Princeton University’s Environmental Institute in New Jersey.

One way round this problem might be to take the carbon dioxide and bury it underground. Another would be to replace fossil-fuel feedstock with biomass. That is in some ways an attractive option — but it is also, as yet, an immature technology.

Expense is another issue. To date, Fischer–



Around the bend: is there a feasible future for liquid fuels made from coal?

Tropsch has always been rather costly, and thus something people normally start to do only when they have no alternatives. Its first major use was during the Second World War, when the blockaded Nazis produced about 90% of their diesel and aviation fuel with the technologies originally developed by Franz Fischer and Hans Tropsch at the Kaiser Wilhelm Institute for Coal Research in 1923. South Africa began liquefying coal in response to apartheid-era sanctions, and in part as a result of its investment back then, continues to derive about 30% of its fuel from liquefied coal.

To make liquid fuel from coal, you first shatter the long hydrocarbon chains into a mixture of hydrogen and carbon monoxide using high temperatures and intense pressure. This is also the first step for the “integrated gasification combined cycle” plants, seen by many as the future of coal-fired generation — a technology that has many synergies with coal-to-liquids.

In Fischer–Tropsch synthesis, the gas is not burned but channelled to a reactor where catalysts reunite the carbon and hydrogen to form hydrocarbon chains of varying lengths, including diesel and petrol. During both phases — gasification and liquefaction — some carbon is given off as carbon dioxide.

Because contaminants such as mercury and sulphur can inhibit the reaction, companies have a built-in incentive to remove impurities from the gas before liquefying it. And the choice of catalyst allows the make-up of synthetic fuel to be tailored to an extent. As a result, diesel produced by the Fischer–Tropsch process is quality stuff. It contains less sulphur and fewer contaminating aromatic compounds, such as benzene and toluene, and releases fewer particulates when burnt than regular diesel fuel does.

Capturing carbon

But none of that solves the carbon dioxide problem. In the United States, all coal-to-liquid plants on the drawing board would include carbon capture followed by, in most cases, sequestration, says Lowell Miller, director of the US Department of Energy Office of Sequestration, Hydrogen, and Clean Coal Fuels. That includes the 22,000-barrel-a-day operation near the town of Roundup that Schweitzer has proposed. But the Natural Resources Defense Council (NRDC), a US-based environmental group, says that even if 90% of that carbon were captured, producing and using coal-derived fuels would still release 8% more carbon dioxide than petroleum-derived fuels¹. “Even if you do carbon sequestration, at best coal-to-liquid methods are still no better than crude oil in terms of lifecycle emissions,” says David Hawkins, director of the NRDC’s Climate Center. “And if we are building a new industry to make transportation fuels, we need to build an industry that produces fuels that are significantly lower in carbon dioxide emissions.”

In China — which, like the United States, is not bound by the Kyoto Protocol, and which has vast coal reserves — carbon sequestration is less likely. Yong-Wang Li, director of Synfuels China in Shanxi, says that there are two proposed coal-to-liquid industrial plants under consideration in China and that, at present, neither proposal contains plans for sequestering carbon.

The advantage of using plant biomass as a feedstock from which to make synthetic fuel, on the other hand, is that no sequestration is necessary — the emitted carbon is carbon that came

W. CAMPBELL/CORBIS



from the air in the first place. If one were to add sequestration to a biomass-to-liquids plant, the result could be 'carbon negative', in that the net effect on the atmosphere would be to draw down the level of carbon dioxide as some of the carbon dioxide fixed by the plant would be sequestered into the planet's crust. What's more, Fischer-Tropsch methods could complement, at the very least, some other biomass technologies.

Plant material that contains too much lignin and not enough cellulose for use in cellulosic ethanol projects (see page 673) could still be used in a Fischer-Tropsch system. The technique could open up the possibility of using forest thinnings and other sources of wood waste, or even the lignin-rich residue left over from cellulosic ethanol production, says David Dayton, a senior scientist at the National Renewable Energy Laboratory in Golden, Colorado. Others think that Fischer-Tropsch could outcompete the fermentation of cellulose more generally.

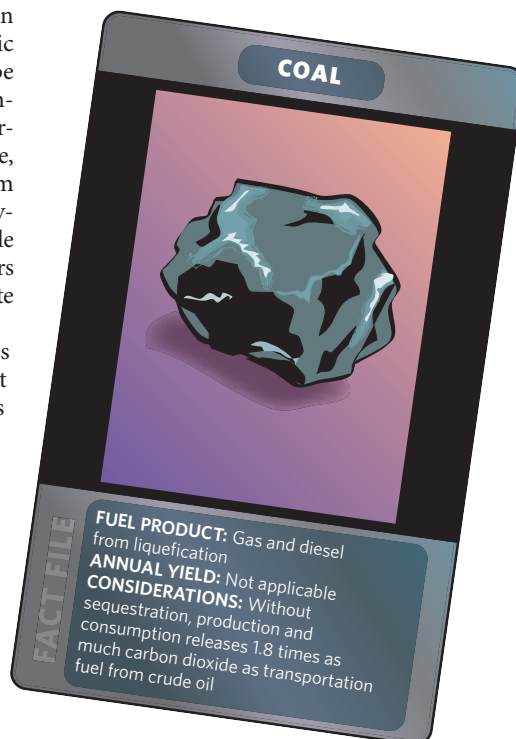
But the technology for gasifying biomass currently lags well behind the development of such methods for coal. The Netherlands has invested a lot of research into the process, but still operates only demonstration-scale biomass-only plants. Choren, a German company, and Shell are building a commercial plant in Freiberg, which will produce 15,000 tonnes per year (110,000 barrels per year) of what Choren calls SunDiesel. Construction of five 200,000-tonnes-per-year facilities that will use wood and agricultural waste is scheduled to begin in 2008. By far the biggest undertaking of its kind to date, the total output from this project

would still be enough to supply only about 4% of Germany's projected diesel needs in 2015.

The other big issue is cost. Several US demonstration coal-to-liquid plants, established during the spike in oil prices during the 1970s, closed without leading to further commercial development after oil prices fell in the early 1980s. Miller and Jim Bartis, a policy analyst at the RAND corporation in Santa Monica, California, both think that oil prices must be above \$50–\$55 a barrel (159 litres) for coal-to-liquid plants to make long-term economic sense. With oil prices currently just under \$60 a barrel that's already an uncomfortably snug fit – and there's no guarantee of prices staying at that level.

Biomass boost

Nevertheless Chevron, Shell and Exxon have all invested in development of Fischer-Tropsch technologies, as has the biggest US coal company, Peabody Coal, which is working on Schweizer's Roundup plant with Rentech of Denver, Colorado. "Some of the big players are willing to take a low rate of return just to establish a technology position," says Bartis. "Once you build a first plant, you're going to learn by doing, and subsequent plants are going to cost less." Another strategy is to concentrate on people who might pay a premium for domestic hydrocarbons. Syn-troleum of Tulsa, Oklahoma, recently provided samples of its natural-gas-derived jet fuel, made with technology licensed from Exxon Mobil, to the US Air Force for testing.



For biomass, the situation is worse. Robin Zwart of the Dutch Energy Research Centre in Petten hopes that upcoming improvements in efficiency will drive the price down, but says that oil prices will still have to exceed \$70–80 per barrel to make liquid fuels from, for example, willow trees economical. That said, carbon taxation or emissions trading would give a boost to biomass-based systems that are unavailable to coal-to-liquid systems.

That is why, until biomass supply and technology are scaled up, there is still the appealing option of spiking coal feedstock with biomass. Coupled with carbon sequestration, this would reduce greenhouse gas emissions without requiring much change to existing technology, says Robert Williams, a researcher at Princeton University's Environmental Institute. Williams has calculated that a mixture of 89% coal and 11% biomass could reduce carbon emissions by 19% relative to using the same process with coal only.

Because they can't yet make money, current Fischer-Tropsch projects often involve a complex mix of industry partners and government subsidies. Sasol of South Africa, a veteran in the easier gas-to-liquids game, considers government subsidies crucial. Sasol is conducting feasibility studies for coal-to-liquid projects in China and India, and has partnered with oil giant Chevron in Europe. But it is waiting to see what develops with government subsidies, according to chief executive Pat Davies, before committing to any US projects.

So far, the US federal government has proposed tax credits for coal-to-liquid programmes, and provides grants to interested companies. States are also pitching in: Pennsylvania, for example, is guaranteeing \$465 million in loans and \$47 million in tax credits for a proposed plant in Schuylkill County. Elsewhere in the world — in China, India and the Philippines, for example — liquefaction projects have received pledges of strong government support. And in Germany, biomass-derived fuels are exempt from the heavy taxes levied on other fuels.

Miller says he is optimistic that Fischer-Tropsch fuels could finally establish a foothold in the United States. But 30 years of studying coal-to-liquid technology has taught him to temper his enthusiasm with caution. "I've been in this business for a long time," he says, "and I'll believe it when the shovel goes into the ground."

Heidi Ledford is a science writer currently in Nature's Washington DC office.

1. Williams, R. H., Larson, E. D. & Jim, H. 8th International Conference on Greenhouse Gas Control Technologies, Trondheim, Norway, 19–22 June 2006.

See Editorial, page 654.

A timely wake-up call as anti-evolutionists publicize their views

SIR — Your Special Report “Anti-evolutionists raise their profile in Europe” (*Nature* **444**, 406–407; 2006) mentions a seminar held in Brussels at the European Parliament on 11 October 2006, as part of a new strategy by supporters of intelligent design (ID) to disseminate anti-evolutionism among the general public of Europe. Two days later, the Catholic Kolbe Center for the Study of Creation and the creationist group Truth in Science published summaries on the Internet. A moderator of the seminar, Maciej Giertych, then published a Correspondence (“Creationism, evolution: nothing has been proved” *Nature* **444**, 265; 2006) claiming that his arguments are entirely scientific and denying any religious component to them. I believe, therefore, that it was a good decision by *Nature* to publish this Correspondence, as a wake-up call to scientists.

The anti-evolution seminar was a series of three public lectures, introduced and moderated by Giertych, who is the retired head of the genetics department of the Polish Academy of Sciences and an honorary member of the Daylight Origins Society, a Catholic creationist organization based in Britain. The seminar was co-organized by Dominique Tassot, director of the Centre d'Etude et de Perspectives sur la Science, an association of 700 Catholic intellectuals who do not accept macroevolution because it is in conflict with their interpretation of the Bible (see *Nature* **439**, 534; 2006).

At the meeting, Giertych pointed out that macroevolution (the gradual appearance of novel body plans as documented in the fossil record) is a “falsified hypothesis” and that there is, from genetics research, no evidence but “only disproof” for Darwin's principle of common descent of all life on Earth. These claims were supplemented by Joseph Mastropaolo, a US aerospace physiologist, who argued that the theory of evolution, after more than 150 years, “still lacked any empirical proof”.

The German civil engineer Hans-Joachim Zillmer told the audience that the fossil record does not provide evidence for gradual macroevolution. Zillmer was announced as an expert in palaeontology and evolution, but he has not, according to the Web of Science, published any paper in the peer-reviewed literature. He is the author of popular books with titles such as *Darwin's Mistake* (Adventures Unlimited Press, 2003) or *Die Evolutionslüge* (*The Evolution Lie* Langen Müller, 2005). In *Darwin's Mistake*, Zillmer asserts that he has found human and dinosaur footprints in fossil-bearing sediments in a riverbed in Texas and concludes that these organisms

lived together. Even creationists no longer claim that these supposed ‘human prints’ are genuine (see *Nature* **323**, 390; 1986). Zillmer's books state that biologists, geologists and the editors of most scientific journals are either misled or fools.

Finally, Guy Berthault told the audience about his research on the rates of sediment depositions, which “did not form slowly over millions of years”, but “have been laid down within very short time periods”. Hence, according to Berthault, most geological data on the age of fossils must be wrong. Giertych's controversial letter is a brief summary of these anti-evolution, pro-ID-lectures.

U. Kutschera

Institute of Biology, University of Kassel,
Heinrich-Plett-Strasse 40,
D-34109 Kassel, Germany

Creationist views have no basis in science

SIR — Maciej Giertych signed his Correspondence letter (*Nature* **444**, 265; 2006) as an employee of the Institute of Dendrology, the Polish Academy of Sciences. As the director of the institute, I would like to point out that, although I respect Professor Giertych's rights to express his views, they are not endorsed by our institute. In my opinion, creationism has no basis in science and should not be regarded as scientific.

Gabriela Lorenc-Plucińska

Institute of Dendrology, Polish Academy of
Sciences, Kórnik, Poland

Creationists weaken society's trust in scientists

SIR — As a scientist I was surprised, and as a Polish scientist I felt ashamed, to read Maciej Giertych's view published in *Nature* (*Nature* **444**, 265; 2006). I would like to assure you that biologists in Poland do follow current scientific findings and would strongly disagree with several statements made in that letter.

There is no accepted scientific evidence for his most ridiculous claims: exclusively harmful mutations, reduction of genetic information or the coexistence of dinosaurs and humans. The only statements I would agree with are that scientists have to search for explanations of what they see in the world around them, and that they should be critical about both new and well established findings.

Polish politicians' recent denial of the theory of evolution is very dangerous, not only because it goes against the scientific paradigm, but also because it weakens society's trust in scientists and in research. Our protests have gained support even from Polish academics with religious connections,

such as Catholic lecturers in the philosophy of nature. The publication of unsubstantiated claims and incorrect statements in renowned scientific journals gives undeserved support to the creationist movement.

Joanna Rutkowska

Institute of Environmental Sciences, Jagiellonian
University, Gronostajowa 7, 30-387 Kraków, Poland

Claim of bias against critics is refuted by publication

SIR — Maciej Giertych states in Correspondence (*Nature* **444**, 265; 2006): “I believe that, as a result of media bias, there seems to be total ignorance of new scientific evidence against the theory of evolution.” However, he does not refer to one publication in a peer-reviewed scientific journal to support the existence of any such “new scientific evidence”; nor has he himself published any. Until any such publication, the existence of scientific evidence against evolution remains unsubstantiated. Further, where is the bias of which Giertych speaks? The very fact that his letter was published shows that *Nature* has no bias against critics of evolution.

Gerdien de Jong*, Gert Korthof†

*Evolutionary Population Biology, Utrecht University,
Padualaan 8, NL-3584 CH Utrecht, the Netherlands
†Bilthoven, the Netherlands

Pseudoscience should not be published in *Nature*

SIR — Although we acknowledge the need to allow publication of diverse opinions in the name of free speech, *Nature* has a responsibility, as a leading and widely read science journal, to uphold scientific standards and values. Unfortunately, in Maciej Giertych's Correspondence letter (“Creationism, evolution: nothing has been proved” *Nature* **444**, 265; 2006), *Nature* fell short in this duty, allowing creationist pseudoscientific arguments to be presented as fact, without any supporting evidence.

The arguments used by Giertych are widely used by creationists, and, in their pseudoscientific tradition, evidence that discredits them is constantly ignored. For example, his suggestion that dinosaurs coexisted with humans, presumably based on supposed human footprints found alongside those of dinosaurs in the Glen Rose Formation of Texas (as expounded by Henry M. Morris in *Scientific Creationism* CLP Publishers, 1974) has been refuted: the ‘human’ footprints are now recognized as dinosaurian (R. Hastings *J. Geol. Educ.* **35**, 4–15; 1987). A comprehensive source that scientifically discredits such ‘evidence’ can be found at <http://scienceblogs.com/>

pharyngula/2006/11/nature_publishes_a_crank_lette.php#more.

We worry that Giertych's Correspondence will lend credibility to pseudoscientific efforts to undermine evolutionary theory. Its publication is damaging to *Nature's* reputation and to science itself. We as scientists may be able to see whether a claim is scientifically thorough, but many other people cannot. We urge the editors to insist on the same scientific rigour in Correspondence as in any other section of *Nature*.

Uwe Balthasar, Susannah Maidment

Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK

There is no new evidence that undermines evolution

SIR — We are astonished that *Nature* would publish a Correspondence as full of errors as that by Maciej Giertych (*Nature* 444, 265; 2006). For someone with degrees from the universities of Oxford and Toronto, Giertych displays a breathtaking ignorance.

There is no “new scientific evidence against the theory of evolution” as he asserts, but fails to document. This can be verified by consulting any of the recent standard textbooks on the subject. The claim that “microevolution... is a step towards a reduction of genetic information” is nonsense. On the contrary, there is ample evidence for the frequent use of duplications of genes in evolution, many of which have acquired new functions. By any criterion, this represents an increase in the amount of genetic information.

Contrary to Giertych's statements, the temporal ordering of rock layers by stratigraphy, and the extinction of dinosaurs some 65 million years before the existence of humans, are overwhelmingly established facts of geology and palaeontology. His claim that “No positive mutations have ever been demonstrated” is simply false. Disregarding the fact that it is illogical to rule out resistance to antibiotics and herbicides as examples of adaptations, as was done by Giertych, there are literally thousands of cases in which natural selection has been demonstrated in wild populations of animals and plants.

Further, the contemporary literature on molecular evolution is filled with studies that provide evidence for a positive role of natural selection. Physicists do not spend their time debating the correctness of the atomic theory of matter; it is intolerable that biologists should constantly be forced to defend their unifying theory against ill-informed attacks.

Brian Charlesworth and 34 others (names available on request from B.C.)

Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3JT, UK

Walking with dinosaurs? Not in the real world

SIR — It is to be hoped that Maciej Giertych's comments in Correspondence (*Nature* 444, 265; 2006) will generate a flood of refutations. Staying within my archaeological profession, Giertych's claims that there are data suggesting that “dinosaurs coexisted with humans” or that there was a “major worldwide catastrophe in historical times” are simply false. These are claims regularly made by religious fundamentalists in support of creationism, exposing Giertych's assertion that his objections are scientific.

In support of these claims, some creationists promote known frauds such as the Paluxy River ‘human footprints’ and the ‘dinosaur figurines’ from Acambaro, Mexico, which they misrepresent as modelled from living observations. There are many creationist interpretations of prehistoric rock art that owe more to Hermann Rorschach than to Richard Owen.

Creationists also distort actual science. Mary H. Schweitzer's research on dinosaur tissue preservation has been used for years as ‘proof’ that Earth is a few thousand years old, as I document in “Dino-blood and the Young Earth” and “Dino Blood Redux” (see www.talkorigins.org). Giertych's Correspondence shows us the irrational basis of creationism in the twenty-first century and warns the international scientific community that this delusion is not restricted to American hillbillies.

Gary S. Hurd

33902 Silver Lantern, Dana Point, California 92629, USA

Creationists pose political, not scientific, threat

SIR — I was disappointed to see Maciej Giertych's letter “Creationism, evolution: nothing has been proved” (*Nature* 444, 265; 2006) published without any disclaimer or comment by the editors. Even though I am aware that *Nature* asked Giertych to comment on the News story “Polish scientists fight creationism” (*Nature* 443, 890–891; 2006), I can find no justification for publishing pseudoscientific arguments in this first-rate scientific magazine.

The level of scientific illiteracy in those arguments is self-evident and, as such, does not need any further discussion. What needs some comment is Giertych's claim about the scientific inspiration for his criticism of evolutionary science. Contemporary creationists espousing ‘intelligent design’ (ID) are careful to avoid mentioning religion — as was Giertych in his recent public statements. Yet Polish readers can refer to his four articles in *Encyklopedia “Białych Plam”*

(*The Encyclopedia of ‘missing pages’* volumes 4 and 6, PWE, 2000, 2001). These articles, “Darwin, Charles Robert”, “Darwinism”, “Evolution”, and “Evolutionism”, provide a more extensive version of the arguments presented in his Correspondence, and explicitly refer to religion and ID views.

Reasonable criticism is as fundamental to science as natural selection is to adaptive evolution. But what Giertych calls “new scientific evidence against the theory of evolution” could not be published in any serious peer-reviewed journal. In fact, publishing scientific papers is not a significant goal for creationists in Poland or anywhere else — on the contrary, the goal is to replace evolution with some pseudoscience in school curricula, as reported in the News story “Polish scientists fight creationism” (*Nature* 443, 890–891; 2006). The creationists' movement is dangerous to the general public on political, not scientific, grounds.

Jerzy Banbura

Department of Experimental Zoology and Evolutionary Biology, University of Lodz, Banacha 12/16, 90-237 Lodz, Poland

How the word ‘hominid’ evolved to include hominin

SIR — Human evolution has long been a subject that can claim a love of tongue-twisting terminology, which if not properly explained can lead to much confusion. The latest example is the reassessment of the linnaean description of the relationship between *Homo sapiens* and the other great apes. To better describe the close evolutionary relationship between them, a new sub-family level has been created. The family group ‘hominid’ now contains all of the African apes, not just the species of the human lineage; and the newly created sub-family name ‘hominin’ (with associated sub-families for *Pan* and *Gorilla*) contains all the species of the human evolutionary lineage. The term hominin is now used where hominid was previously, causing much confusion, especially among students and nonspecialists.

Publications have a key role in ensuring that this change and the reasons for it are as clear as possible. Yet journals, including *Nature*, continue to publish papers and features that use the terms hominin and hominid interchangeably. If all in our discipline could agree to the new terminology, journals could ensure that it is used correctly, bringing a small amount of clarity to a subject that so often presents problems of interpretation for the nonspecialist but fascinated public audience.

Simon Underdown

Department of Anthropology and Geography, Oxford Brookes University, Gipsy Lane Campus, Headington, Oxford OX3 0BP, UK

COMMENTARY

The right tools can save lives

Effective diagnosis, paired with treatment, for developing-world diseases can have far-reaching impacts, says the **Global Health Diagnostics Forum**.

Millions of people in developing countries die each year from diseases that are treatable or preventable, and just three diseases — AIDS, tuberculosis (TB) and malaria — kill more than 5 million people annually¹. Effective treatment for these and other conditions is limited by the failure to accurately diagnose disease in resource-limited settings. For example, the standard diagnostic test for TB, which is latently carried by one-third of the human population and kills roughly 2 million people each year, misses half of all cases².

Diagnostic technologies are needed not just for identifying disease, but also for designating the appropriate treatment, monitoring the effects of interventions (preventive or therapeutic) and for determining drug resistance. Effective diagnostics, paired with access to treatment, can have a huge impact on health.

Most diagnostic tools were developed for use in industrialized countries, and thus are too complex and expensive for developing-world conditions. All too often such tools are not evaluated in relevant settings³, and suffer from undertrained end users, insufficient advocacy and inadequate quality assurance⁴.

Forum formed

Despite the severe shortage of effective developing-world diagnostics, little has been done to address the problem. This is due to a scarcity of information on the potential health impact and performance of essential diagnostics, and to the low return on investment in diagnostics perceived by the private sector.

To provide information on specific diagnostic needs and characteristics, the Bill & Melinda Gates Foundation, in partnership with the RAND Corporation, convened the Global Health Diagnostics Forum with representatives from the diagnostics industry, technology developers and experts in relevant diseases and public health. Between 2004 and 2006 the Forum focused on six disease groups that have the highest health tolls in the developing world: acute lower respiratory infections (ALRIs), HIV/AIDS, diarrhoeal diseases, malaria, TB and sexually transmitted infections.

For each disease, the Forum identified one or two points along the path of disease progression where a diagnostic test could have the most impact, and then used models to determine the health impact and the required diagnostic characteristics. A key factor was

the eventual healthcare setting: these ranged from severely deficient infrastructure common in home settings (for example, no water, electricity or staff expertise), to moderate or advanced infrastructure found in hospitals and some Asian urban clinics.

The Forum also determined the potential health impact of a specific diagnostic tool for each disease⁵, assuming that access to diagnostics is linked to effective interventions.

The big six

A diagnostic test for bacterial ALRIs could save many lives if it reduced unnecessary antibiotic use while encouraging more appropriate use. A new diagnostic test with 95% sensitivity and 85% specificity, accompanied by greater treatment access and minimal infrastructure requirements, could save at least 400,000 adjusted lives (which also account for lives saved due to reductions in over-treatment) a year. A test for severe cases of ALRI requiring hospitalization would also be beneficial, if access to effective hospital care increases globally.

A test for paediatric HIV infection in infants younger than 12 months with 90% sensitivity, 90% specificity, and minimal laboratory requirements, could save about 180,000 disability-adjusted life years (DALYs) over 12 months if 5% of the target population has access to antiretroviral therapy, or around 2.5 million DALYs if 100% has access.

A test for diarrhoeal diseases with 90% sensitivity, 90% specificity, and minimal infrastructure requirements for each of the pathogens *Giardia lamblia*, *Cryptosporidium parvum* and enteroaggregative *Escherichia coli* could reduce the prevalence of stunted growth by 12.5% and save 2.8 million DALYs annually. But this finding depends critically on uncertain assumptions about treatment cost and additional health benefits (such as an improved immune response), both of which require further research.

A paediatric test for malaria with 95% sensitivity, 95% specificity, and minimal infrastructure requirements, could avert more than 100,000 childhood deaths and around 400 million unnecessary treatments annually. A test with no infrastructure requirements

and 90% sensitivity and specificity could avert more than 300,000 childhood deaths and some 450 million unnecessary treatments annually.

For syphilis, a test that is at least 86% sensitive, 72% specific, requires minimal infrastructure, and has either 100% rate-of-return for test results or 100% treatment rate could save at least 138,000 adjusted lives and avert more than 148,000 stillbirths annually. A similar test requiring no laboratory infrastructure could save more than 200,000 adjusted lives and avert 215,000 stillbirths per year.

A diagnostic that requires minimal infrastructure and has 85% sensitivity and 90% specificity for both gonorrhoea and chlamydia could save nearly 3 million DALYs, avert more than

12 million infections, and prevent at least 161,000 related HIV infections among female sex workers in sub-Saharan Africa, China and South East Asia, over 4 years.

A rapid TB diagnostic requiring no laboratory infrastructure that has 97%

specificity, at least 85% sensitivity, and is unaffected by HIV status, could save around 400,000 adjusted lives annually.

In addition to the above tests, diagnostics that can identify and distinguish between multiple diseases with similar symptoms (for example, the acutely ill child who presents with fever) would improve health outcomes.

The Forum has defined the need and impact of diagnostics for six devastating disease groups. Now we challenge scientists, technology developers, funding agencies, policy-makers, international governmental and aid organizations, investors and diagnostic companies to work together to take this forward. A coordinated approach is needed so that appropriate diagnostics can achieve the promised impact. ■

A full list of signatories to this Commentary is available as supplementary information at www.nature.com/nature/journal/v444/n7120/supinfo/444681a.html.

e-mail: diagnostics@gatesfoundation.org

"The standard diagnostic test for TB, which kills 2 million people each year, misses half of all cases."

1. Lopez, A. D. et al. *Global Burden of Disease and Risk Factors* (Oxford Univ. Press, 2006).
2. WHO *Global Tuberculosis Control* (Geneva, 2004).
3. Mabey, D. et al. *Nature Rev. Microbiol.* **2**, 231–240 (2004).
4. Lon, C. T. et al. *Trans. R. Soc. Trop. Med. Hyg.* **99**, 493–498 (2005).
5. Burgess, D. C. et al. *Nature Suppl.* S1 (2006).

BOOKS & ARTS

One good deed

Can a simple equation explain the development of altruism?

The Altruism Equation: Seven Scientists Search for the Origins of Goodness

by Lee Alan Dugatkin

Princeton University Press: 2006. 224 pp.
\$24.95, £15.95

Oliver Curry

The evolutionary explanation of altruism will surely come to be seen as one of the major breakthroughs of twentieth-century science. At first, the problem of altruism seemed insurmountable: how could natural selection favour the tendency to help others? Surely the ruthless, competitive process of evolution could give rise only to ruthless, competitive individuals. Altruism, morality and goodness must come from somewhere other than biology.

Enter 'selfish gene' theory — the recognition that genes, not individuals, are the units of selection. This at once both clarified and helped to solve the problem. It became clear that although genes must be 'selfish', in the sense of promoting their own replication, individuals need not be. On the contrary, genes can sometimes get themselves replicated by building social, cooperative and altruistic individuals.

How? Well, genes for altruism can spread if they help copies of themselves that reside in other individuals — hence that oasis of altruism we call the family (this is kin altruism). They can spread if they build organisms that team up with others to form mutually beneficial partnerships (mutualism), or that exchange benefits over time (reciprocal altruism). They can spread by building organisms that settle disputes over status and resources not by fighting but by conceding to prior ownership, or by deferring to the victors of ritual contests (conflict resolution). And genes for altruism can spread if altruism serves to attract mates (sexual selection).

These theoretical advances led to an avalanche of empirical work on the biological basis of social behaviour; this was 'sociobiology', and it laid the foundations for the scientific investigation of human behaviour. In *The Altruism Equation*, Lee Alan Dugatkin tells the story of this scientific revolution through mini-biographies of the chief protagonists.

We learn that Charles Darwin regarded altruism — in the form of sterile castes of worker bees that forgo their own reproduction



What's in it for the workers? Altruism in bees initially puzzled Darwin (inset).



and sacrifice their lives for the hive — as a "special difficulty, which at first appeared to me to be insuperable, and actually fatal to the whole theory". But we also discover he had an inkling that the problem might be solved by 'family selection'. We learn that T. H. Huxley regarded the family as the sole haven of altruism, and viewed the rest of nature as a bloody, gladiatorial arena. We learn that Petr Kropotkin thought that evolution penalized individuals who fought each other, but that it favoured those who combined to combat harsh environments. We learn that each of the triumvirate of the modern synthesis — J. B. S. Haldane, Ronald Fisher and Sewall Wright — came tantalizingly close to cracking the secret of kin altruism but, inexplicably, neglected to formulate their insights mathematically. We learn that W. D. Hamilton — an alumnus of my own institute, the London School of Economics — toiled alone to arrive at the $E=mc^2$ of evolutionary biology, $rB>C$ (where r is the coefficient of relatedness, B is the benefit and C is the cost). And we see how his elegant theory sheds light on phenomena as diverse as coloration in butterflies, helpers at the nest, alarm calls in ground squirrels, and parental care in human stepfamilies.

These biographies help to put the theories in their historical context, and bring warmth and colour to the story. Unfortunately, however,

this device can also obscure the science.

First, *The Altruism Equation* focuses on kin altruism at the expense of other theories of cooperation. The theory of mutualism is never fully explained; reciprocal altruism, which is thought to be responsible for trust, gratitude, guilt and revenge in humans, gets half a chapter; ritual contests get just half a page; and the recognition of property, costly signalling and sexual selection are not mentioned at all, even in the index. There is no discussion of conflict and cooperation between genes in the same individual, the evolution of fairness, coalition formation, or recent research on cooperation in humans. This is all the more puzzling because Dugatkin's previous work shows that he is familiar with, and has indeed contributed to, the development of these theories.

Second, the logic of kin selection struggles to be heard over the din of its imprecise historical precursors. For example, kin selection depends on the chance of a particular gene for altruism being present in another individual, and not merely with the overall proportion of genes that individuals share. But readers could easily miss this point. The logic is further obfuscated by Dugatkin's folksy way of referring to kin altruism as occurring between 'blood relatives', rather than genetic relatives. This is particularly unfortunate because kin altruism can occur between organisms that have no blood at all, such as plants and bacteria.

Third, some of the early players in the story contribute little to the modern theory, with the result that the chapters devoted to them are somewhat repetitive. So, for example, Kropotkin and Huxley's 'good cop–bad cop' routine is needlessly iterated, and we are told of W. C. Allee's interest in establishing the notion of altruism as divorced from kinship more than 25 times in a 23-page chapter.

Still, you can't keep a good theory down. And if the book inspires readers to explore the area further, it will have been a success. ■ Oliver Curry is at the Centre for Philosophy of Natural and Social Science, London School of Economics, Houghton Street, London WC2A 2AE, UK.

K. TAYLOR/NATUREPL.COM

NATIONAL PORTRAIT GALLERY, LONDON/BRIDGEMAN ART LIBRARY

FILM

The quest for immortality

The Fountain

directed by Darren Aronofsky
Warner Brothers
US release, 22 November 2006

Emma Marris

The Fountain defies characterization by genre. Aimed at arthouse audiences, it uses one actor — Hugh Jackman — to play the lead in three ambiguously related stories. Is he the same man? We're not sure. The three stories are set hundreds of years apart, but the common theme is the search for immortality, so he could be. The film won the Alfred P. Sloan Foundation's annual prize for a feature film dealing with science and technology at the Hamptons International Film Festival this October. Its director, Darren Aronofsky, has previously directed *Requiem for a Dream* and π , a film with a mathematician as protagonist.

In one of *The Fountain*'s three strands, a scientist (Jackman) races to find a drug that will stop the growth of brain tumours before his wife (Rachel Weisz) dies of one. In dramatizing this situation, Aronofsky compresses into about four minutes the whole process of drug discovery, from lead identification right through primate testing. In scenes that intentionally call to mind TV hospital dramas, lab-coated scientists whirl around barking jargon at each other, their eyes wide with earnest concentration.

Here, a researcher in the audience might think, is science finally presented as the dramatic and compelling endeavour it is. The film makes science look sexy, and without wholly departing from actual lab realities. Indeed, the film was co-written by Ari Handel, who has a



WARNER BROS PICTURES

In a biosphere bubble in the future, Hugh Jackman's character in *The Fountain* accepts the idea of death.

PhD in neurology but left the academic track to make films with his college room-mate, Aronofsky. Handel's job was to keep an eye on accuracy. At one point, the hero's boss snaps at him for testing a drug on a monkey on a whim, saying: "The NIH could shut us down." On the other hand, the lab in the film is ridiculously stylish and tidy, but then Hollywood always gives characters apartments they couldn't possibly afford too.

The film is not really about science, but rather attitudes to death. In the first strand, Jackman plays a Spanish conquistador who hopes to use Christianity to cheat death, which is wrong, wrong, wrong, the film says. And it

is the Jackman character in the third segment who, while flying through space in an attractive biosphere bubble and working on his weightless tai chi, figures out that the appropriate response to death is acceptance. For those who are impatient with modern, diffuse spirituality, this comes across as pretty silly stuff.

Back in the central science strand, Jackman's character says: "Death is a disease just like any other, and there is a cure and I will find it." But here we are seeing one of the oldest tropes about scientists played out yet again: the scientist as an allegorical figure of hubris.

Emma Marris is a *Nature* reporter based in Washington DC.

Bugs with bugs

Big Fleas Have Little Fleas: How Discoveries of Invertebrate Diseases are Advancing Modern Science

by Elizabeth W. Davidson

University of Arizona Press: 2006. 208 pp.
\$35 (hbk); \$17.95 (pbk)

Mark L. Winston

Insects get sick too, and the curious invertebrate pathologists who have delved into this diseased world have revealed some compelling insights into nature, illness and scientific practice. That is the basic premise of *Big Fleas Have Little Fleas* by Elizabeth Davidson, a book based on tales about the diverse organisms that plague not just insects, but crustaceans and horseshoe crabs too. The author

draws on these pathologies to open the door on biological complexity and the splendour of scientific enquiry.

Davidson takes us through stories of infestations and plagues, uncovering how scientists have methodically unravelled both basic principles and potential control measures by studying parasites and diseases of invertebrates. European investigations into rotting silkworms in the nineteenth century found maladies caused by fungi and protozoans, and inspired Pasteur to establish the field of epidemiology. Elegant research into cholera found the causative bacteria *Vibrio cholerae* hitchhiking globally on copepods, and led US scientist Rita Colwell to suggest filtering drinking water in India through old sari cloth, a simple

control measure that cut infection rates in half. Infestations of gypsy moths today are partly controlled by an arsenal of pathogens whose biology and utility have emerged from more than a hundred years of research.

Much of the work described in the book occurred during the founding era of invertebrate pathology, before the availability of technologies such as gas chromatography, electron microscopy and gene sequencing. Davidson's book reminds us that intuition, rigorous thinking and thorough probing were the most important scientific tools for these early researchers.

Unfortunately, Davidson does not use a similar rigour to probe the relevant policy issues, such as why biological control remains a minuscule sideshow compared with chemical pesticides. Viral, fungal, protozoan, nematode and bacterial agents can provide a solution but

are dwarfed by the almost 2 million tonnes of chemicals used globally each year.

Even *Bacillus thuringiensis*, the most widespread biological control agent, is usually used by incorporating its toxins into genetically modified crops, rather than through more ecologically compatible whole-organism field sprays. Rampant antibiotic resistance among human diseases and today's rapidly evolving pandemic agents that afflict or threaten humans, plants and animals are other serious concerns that receive insufficient attention from Davidson.

Big Fleas Have Little Fleas is a book in search of a voice: it is not detailed enough for academic specialists, and not sufficiently well written for a general audience. Chapters start with interesting tidbits, but the writing quality is not sustained. It is a frustrating tease, with occasional elegant moments linked by formulaic descriptions of how this scientist did this, then the next one did that, with each piece of research contributing in a tiny way to larger scientific principles. Yes, science is a slow, methodical and painstaking process, but it's the rare moments of brilliance and the great

investigative stories that thrill readers — elements that are too sparsely described here to make the book a compelling read.

Still, Davidson's book reminds us of one fundamental point: there is still much to learn from contemplating creatures smaller than ourselves, and we have barely begun to unravel the vast biological complexity on which we humans rely. ■

Mark L. Winston is at the Morris J. Wosk Centre for Dialogue, and the Department of Biological Sciences, Simon Fraser University, Vancouver, British Columbia, V6B 5K3, Canada.

The birth of contraception

Contraception and Abortion from the Ancient World to the Renaissance

by John M. Riddle

Harvard University Press: 1992

Michel Raymond

The contraceptive pill — what a wonderful invention! At last, we could have a fulfilling sex life, free from worry about the mischief wreaked by uncontrollable gametes, and could separate the desires for pleasure and reproduction. Let's spare a thought for our poor ancestors, who were faced with the choice of reproducing like rabbits or miserably limiting their sex lives. We've made real progress since then.

Or at least that's what people thought when science delivered modern contraception in the twentieth century. For some reason, this myth — and it is one — still holds, 14 years after the publication of John Riddle's book *Contraception and Abortion from the Ancient World to the Renaissance*.

Birth control, by contraception and abortion, has a long history. In the ancient world there were precise recipes, as we know from books written by the doctors of the time (Soranus, Dioscorides and Hippocrates). These doctors obtained their knowledge from direct contact with ordinary people. One plant in particular was said to be a contraceptive — a giant fennel, *Ferula historica*. It was so sought after and harvested in such quantities that it became extinct.

In the Middle Ages, when universities started teaching medicine, knowledge began to be passed from doctors to their pupils, the next generation of doctors. Contraception was conspicuous by its absence in these courses for men, and such knowledge was lost among doctors. However, it was still transmitted between women — at least while the



Flower power: before the Pill, plants such as the gentian were traditionally used as contraceptives.

traditional way of life continued.

I visited an old alpine village this year that had continued to use traditional agricultural practices until about 20 years ago. An old peasant of 92 told me about a plant with potent contraceptive properties — knowledge she had obtained from her grandmother, who must have learnt about it from her family. The plant concerned was a kind of juniper, the key part being the berries. According to Riddle's book, juniper (which has 23 entries in the index) has been used in contraceptive recipes since ancient times.

The common name of one species of juniper, the savin (*Juniperus sabina*), was derived from its ability to save young women from shame, and modern science has finally confirmed its contraceptive effects. Many of the plants mentioned in old books have had their contraceptive properties confirmed — most of them contain oestrogen.

This traditional knowledge, traces of which remain in the memories of some Europeans, started to disappear with depopulation of the countryside in the nineteenth century: in towns, ancient knowledge ceased to be transmitted. Probably for the first time since the Graeco-Roman era (at least), most Western women no longer had access to an effective means of contraception. The contribution of modern medicine, culminating in the pill, therefore constituted real progress, but it must be seen in the context of history.

Riddle shows us that ancient contraceptive medical practices were safe, effective and commonly used. Sociological studies on their use remain to be carried out. But it is possible that, between the Middle

Ages and the rise of modern contraception, the well-off and city dwellers had little access to effective contraception, thanks to the growth of conventional medicine and the soaring social power of the physician.

This is just one of the many intriguing lines of investigation to arise from this book, which shines a different light on what we are generally taught about the 'progress' of the modern world.

Michel Raymond is at the Institute of Evolutionary Sciences, University of Montpellier II, Montpellier 34095, France.

CANCER

Stem cells and brain tumours

Peter B. Dirks

Stem cells are increasingly implicated in maintaining certain cancers. Studies of an intractable type of brain tumour provide hints as to why such cells may underlie the tumours' resistance to therapy.

Cancers are notorious for their ability to survive treatment and recur. Hopes of understanding how they can do so, however, have grown with the prospective identification of rare populations of cancer stem cells in solid tumours^{1,2}. Two studies in this issue^{3,4} mark a step towards realizing these hopes, and provide further insight into the stem-cell nature of human glioblastoma, an especially nasty type of brain cancer. Both studies build on the identification² of a tumour-initiating subpopulation of cells that express a cell-surface marker, CD133, that is a hallmark of neural precursor cells.

On page 756, Bao *et al.*³ show that glioblastoma cells expressing CD133 (CD133⁺ cells) are resistant to ionizing radiation because they are more efficient at inducing the repair of damaged DNA than is the bulk of the tumour cells. Radiation therapy has been the mainstay of glioblastoma treatment for more than 40 years, but although it is transiently effective, it offers no lasting cure. The implication of these results³ is that radiation treatment fails in the long run because it cannot kill the subpopulation of CD133⁺ tumour-initiating cells.

On page 761, Piccirillo *et al.*⁴ describe their work with bone morphogenetic proteins (BMPs), soluble factors that normally induce neural precursor cells to differentiate into mature astrocytes — a subtype of brain cells called glial cells. The authors show that BMPs can also prompt the differentiation of CD133⁺ brain tumour cells, critically weakening their tumour-forming ability. The results further imply that tumour populations at least partially retain a developmental hierarchy based on stem cells, and remain able to respond to the normal signals that induce them to mature. These findings should lead to renewed interest in devising therapies that promote the differentiation of cancer cells.

Both groups^{3,4} arrived at their findings by considering the functional hierarchy of the heterogeneous population of tumour cells. In doing so, they add weight to the importance in this research of dissociating solid-tumour samples into single-cell suspensions, purifying the stem-cell fractions, and testing their response to treatment. Crucially, both groups

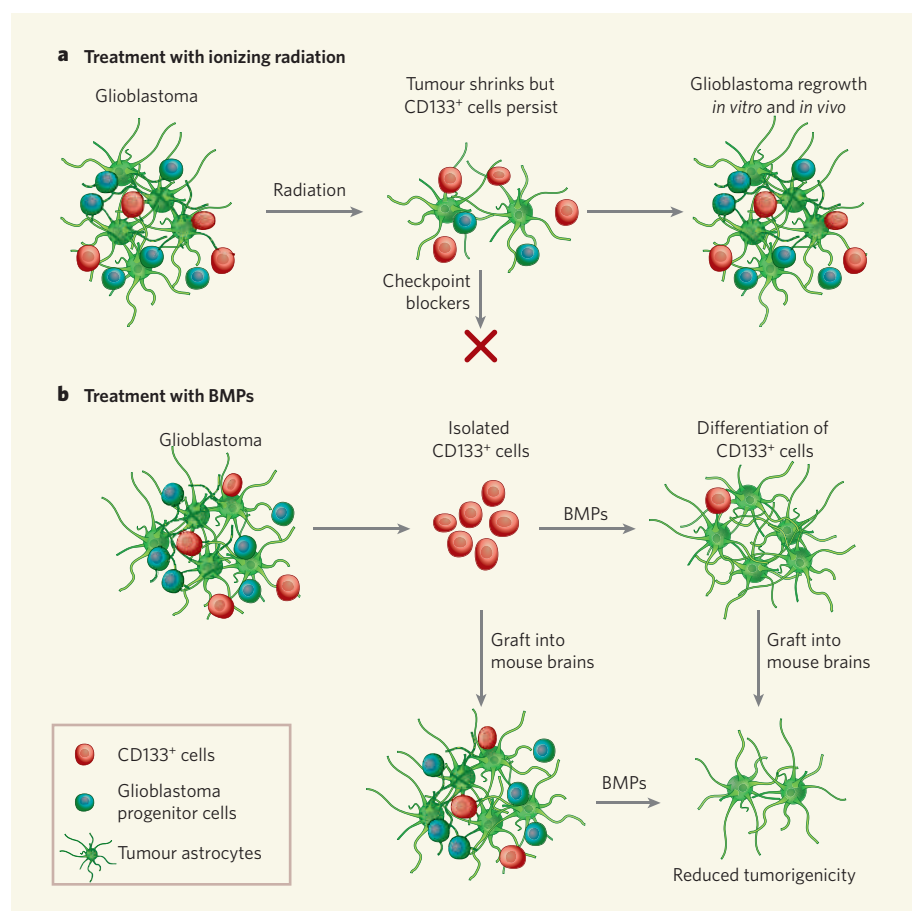


Figure 1 | Response of glioblastomas to ionizing radiation³ and bone morphogenetic proteins (BMPs)⁴. Glioblastomas are heterogeneous tumours that contain a few tumour-initiating CD133⁺ stem cells among other, more differentiated, CD133⁻ cells, including glioblastoma progenitor cells. **a**, Following radiation, the bulk glioblastoma responds and the tumour shrinks. But CD133⁺ cells activate checkpoint controls for DNA repair more strongly than CD133⁻ cells, resist radiation and prompt the tumour to regrow. These cells could be targeted with DNA-checkpoint blockers to render them radiosensitive. **b**, BMPs normally cause neural stem cells to differentiate into astrocytes. When used to treat isolated glioblastoma CD133⁺ cells, they weaken the cells' tumorigenicity both *in vitro* and, when engrafted into mice, *in vivo*. The knowledge that a tumour retains a developmental hierarchy suggests that targeting different cell populations is a promising therapeutic strategy.

verified their *in vitro* results with *in vivo* studies — a true demonstration that human tumour-initiating cells can act as such requires use of the 'gold standard' assay⁵ of transplanting them into immunodeficient mice to see if they retain their stem-cell capacity.

In Bao and colleagues' work³ (Fig. 1a), ion-

izing-radiation treatment of glioblastoma cells grown *in vitro* or as grafts in mice produced an increased fraction of CD133⁺ cells in the residual tumour population compared with that in unirradiated tumours. These cells retained the ability to reinitiate heterogeneous tumours when transplanted into other mice,

demonstrating the retention of stem-cell ability. Although both CD133⁺ and CD133⁻ cells sustained a similar amount of DNA damage, activation of DNA-repair responses was greater in CD133⁺ cells.

Bao *et al.* went further, to foreshadow how this finding might be translated into therapy. They found that CD133⁺ cells could be rendered less resistant to radiation if two agents — the checkpoint kinases Chk1 and Chk2, which control pauses in the cell cycle to allow DNA repair to take place — were inhibited pharmacologically. The authors did not look at whether these cells lost the ability to subsequently initiate tumours *in vivo*. But their work, together with studies showing that leukaemia stem cells are resistant to chemotherapy^{6,7}, means that the idea that cancer stem cells are important because they are resistant to therapy now carries greater weight.

In their research, Piccirillo and colleagues⁴ first showed that human glioblastoma cells express BMPs and their cell-surface receptors. Using culture conditions that support the growth of undifferentiated glioblastoma cells, they found that BMP treatment reduced cell proliferation and induced differentiation predominantly into cells resembling mature astrocytes. Treatment of cultured glioblastoma progenitor cells or CD133⁺ glioblastoma cells *in vitro* with BMP, or co-treatment of glioblastoma cells transplanted with beads soaked in BMP, reduced the size of the tumours grafted into mice and prolonged the animals' survival (Fig. 1b). The BMP-treated tumour cells engrafted into mice were more mature and less invasive. CD133⁺ cells could not be recovered from these small tumours, and such tumour populations were not capable of serial engraftment.

Although these results⁴ are remarkable in showing that a differentiation-promoting agent is a potential treatment for brain tumours, some mice still developed tumours and died 3 months after BMP treatment. So it seems that certain cancer cells — presumably cancer stem cells — still escape BMP treatment. Will these cancer stem cells induce recurrence at a longer latency?

In both cases^{3,4}, definitive demonstration of a specific effect of treatment on glioblastoma stem cells will require improved purification of the tumours — the true stem cells are probably a subpopulation of the CD133⁺ fraction. But both studies add depth to the cancer-stem-cell hypothesis and illustrate the potential of re-examining cancer in the light of that hypothesis. Given the identification of CD133⁺ tumour-initiating cells from human colon cancer, published last month^{8,9}, it would seem that these approaches are ripe for testing in other human cancers. ■

Peter B. Dirks is at the Hospital for Sick Children, University of Toronto, Toronto, Ontario, M5G 1X8 Canada.
e-mail: peter.dirks@sickkids.ca

1. Al-Hajj, M. *et al.* *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).
2. Singh, S. K. *et al.* *Nature* **432**, 396–401 (2004).
3. Bao, S. *et al.* *Nature* **444**, 756–760 (2006).
4. Piccirillo, S. G. M. *et al.* *Nature* **444**, 761–765 (2006).
5. Bonnet, D. & Dick, J. E. *Nature Med.* **3**, 730–737 (1997).
6. Graham, S. M. *et al.* *Blood* **99**, 319–325 (2002).
7. Guzman, M. L. *et al.* *Blood* **105**, 4163–4169 (2005).
8. O'Brien, C. A., Pollett, A., Gallinger, S. & Dick, J. E. *Nature* doi:10.1038/nature05372 (2006).
9. Ricci-Vitiani, L. *et al.* *Nature* doi:10.1038/nature05384 (2006).

PLASMA PHYSICS

On the node of a wave

Tom Katsouleas

A compact electron accelerator can be made by the cunning use of laser pulses to let electrons 'surf' on a plasma wave. The problem has been controlling exactly how much the electrons are accelerated.

In the early 1990s, off the north shore of Maui, Hawaii, Laird Hamilton and Buzzy Kerbox invented the sport of tow-in surfing, using jet skis to propel themselves into particularly large or fast waves. Catching these waves by paddling would have been impossible without becoming caught up in the waves' white water (Fig. 1). On page 737 of this issue, Faure and colleagues¹ describe a similar technique to inject electrons into a plasma wave created in the wake of a propagating laser pulse. The end result is a compact electron accelerator of astonishing stability that could one day find applications in fields from radiation therapy to radiography and femtosecond chemistry.

The authors' result is an improvement on the principle of the 'laser wakefield accelerator', which was used in 2004 to produce electron beams of a single defined energy^{2–4}. The laser wakefield accelerator works by first using a single intense laser pulse to ionize a passive gas such as helium, forming a plasma of charged electrons and ions. Much in the manner of a speedboat passing through water, this pulse creates a wake as it displaces plasma electrons from its path. The wake can gain such a large amplitude that it 'breaks', forming the plasma equivalent of white water⁵ — electrons that

move with the wave, rather than just supporting it as it propagates. The large electric forces associated with the wake capture white-water electrons and accelerate them to energies of up to 200 megaelectronvolts (MeV) over a distance of just 2 mm. That distance is nearly 1,000 times shorter than the length of a comparable conventional accelerator driven by radio-frequency waves. Remarkably, the resulting electron beams are almost monoenergetic. They are also of high quality in the sense that their emittance, a quantity associated with a beam's angular divergence, is small.

Until the latest work, however, the beams produced by a laser wakefield accelerator were not stable: small changes in laser or gas conditions could lead to variations in the final beam energy of as much as 50%. Faure and colleagues' advance¹ is to stabilize the energy gain of the accelerator by introducing an equivalent of a jet ski for the electrons. This takes the form of a second laser beam that controls precisely the way the electrons 'surf' the plasma wave. The result is a high-quality electron beam with a single, well-defined energy that can be varied from 50 to 250 MeV.

To achieve this result, the authors first adjusted the parameters of the plasma and



Figure 1 | Surfing, tow-in style. Pete Cabrinha takes a record-breaking ride on the 70-foot 'Jaws' wave off the north shore of Maui, Hawaii, on 10 January 2004, for which he won the coveted Billabong XXL.

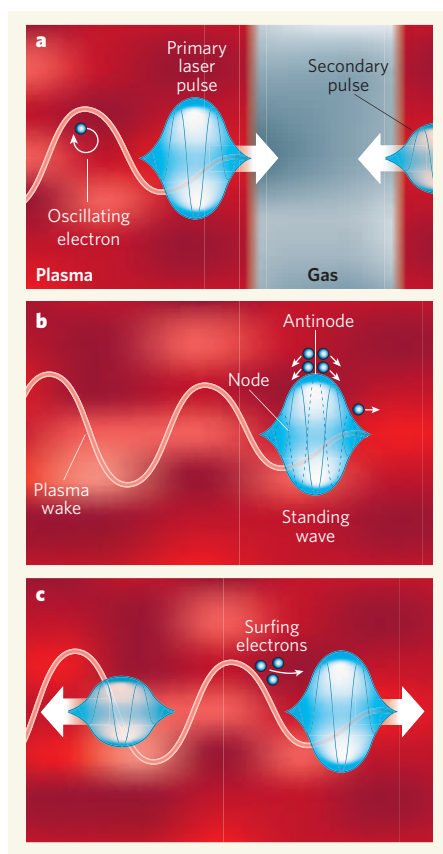


Figure 2 | Surfing in a plasma wakefield. **a**, In Faure and colleagues' experimental scheme¹, the primary laser wakefield pulse ionizes helium gas to a plasma. If the parameters of pulse and plasma are chosen appropriately, the electrons of the plasma oscillate about a fixed spot. **b**, If a second 'tow-in' pulse travelling in the opposite direction crosses the first, a standing wave forms. Electrons are pushed left and right in the standing wave from the antinodes to the nodes. **c**, Some electrons — those pushed to the right — gain enough speed to get caught up in the following wave crest and are accelerated forwards. The energy gain of the electrons is determined by how far they have to surf through the plasma, and so by where exactly along the plasma the two laser pulses cross.

initial laser pulse to keep the amplitude of the plasma waves just below the threshold for wave breaking. This means that no plasma electrons move fast enough to ride the waves. Instead, they just oscillate back and forth, as do water molecules in an ocean wave far from the beach. But when a second pulse is added, propagating in the opposite direction to the first, a temporary standing wave is formed with fixed maxima and minima, known as antinodes and nodes (Fig. 2). The radiation pressure caused by the momentum of the light pushes electrons at the antinodes of the standing wave towards the nodes. This slight extra push is enough to force some electrons into the wake of the first laser, where they become surfers at nearly the speed of light. The location of the overlap of the two lasers determines where along the plasma axis electrons first catch the wave, so the length of their ride, and hence their energy gain, can

be easily controlled by adjusting the timing between the two lasers.

This idea of using a second laser pulse to slingshot electrons into a plasma wake dates back to 1996 (ref. 6) — not long after Hamilton and Kerbox were experimenting with jet skis to sling themselves into ocean waves. Several groups have attempted variations on the original idea, but Faure *et al.*¹ are the first to demonstrate high-quality beam injection. Interestingly, they succeed with the simplest of the schemes proposed so far.

The first idea was to cross two lasers at 90° and use the transverse radiation pressure of the second laser to inject electrons into the main laser's wake. But simulations indicated that it was the second laser's wake, rather than its radiation pressure, that caused injection⁷. The simultaneous use of counter-propagating and co-propagating secondary lasers for injection was then proposed⁸. It was thought that two lasers of frequencies differing by an amount matched to the resonant frequency of electrons in the plasma would be needed to create a slow plasma wave that would push electrons into the main wake. But further simulations⁹ seemed to imply that radiation pressure from the standing field pattern produced with a single counter-propagating secondary laser could be sufficient. It is indeed this last and simplest scheme that has worked to such positive effect.

The approach does have limitations. The charge that can be accelerated in a single bunch amounts to just tens of picocoulombs, which is ten times less than the quantities typically accelerated without a stabilizing second laser. Although this much charge is adequate for many proposed applications^{10–12}, other groups in Britain, the United States and Japan are working with considerable success to stabilize the energy gain at higher charge levels

using more precise control of the laser and gas conditions.

The rapidity with which barriers have continued to fall in the two years since the breakthrough^{2–4} for monoenergetic acceleration of electrons in laser wakefields is astounding. Monoenergetic beams of ions have now been produced by laser irradiation of specially designed foil targets^{13,14}, rather than of gaseous targets as used for electron production. Work is also progressing towards showing that the plasma mechanisms can be scaled to higher energies^{15,16}, and particularly to the stringent requirements of a high-energy collider.

That energy frontier is still a long way off for laser-plasma technology. But thanks to work such as that of Faure and colleagues¹, the widespread use of compact plasma accelerators for medicine, industry and research may be much closer than we think.

Tom Katsouleas is in the Department of Electrical Engineering, University of Southern California, 3737 Watt Way, Los Angeles, California 90089-0271, USA.

e-mail: katsoule@usc.edu

1. Faure, J. *et al.* *Nature* **444**, 737–739 (2006).
2. Mangles, S. *et al.* *Nature* **431**, 535–538 (2004).
3. Geddes, C. G. R. *Nature* **431**, 538–541 (2004).
4. Faure, J. *et al.* *Nature* **431**, 541–544 (2004).
5. Dawson, J. M. *Phys. Rev.* **113**, 383–387 (1959).
6. Umstadter, D., Kim, J.-K. & Dodd, E. *Phys. Rev. Lett.* **76**, 2073–2076 (1996).
7. Hemker, R. G., Tzeng, K.-C., Mori, W. B., Clayton, C. E. & Katsouleas, T. *Phys. Rev. E* **57**, 5920–5928 (1998).
8. Esarey, E., Hubbard, R. F., Leemans, W. P., Ting, A. & Sprangle, P. *Phys. Rev. Lett.* **79**, 2682–2685 (1997).
9. Fubiani, G., Esarey, E., Schroeder, C. B. & Leemans, W. P. *Phys. Rev. E* **70**, 016402 (2004).
10. Brozek-Pluska, B., Gliger, D., Hallou, A., Malka, V. & Gauduel, Y. A. *Radiat. Chem.* **72**, 149–159 (2005).
11. Glinec, Y. *et al.* *Med. Phys.* **33**, 155–162 (2006).
12. Glinec, Y. *et al.* *Phys. Rev. Lett.* **94**, 025003 (2005).
13. Hegelich, B. M. *et al.* *Nature* **439**, 441–444 (2006).
14. Schwöerer, H. *et al.* *Nature* **439**, 445–448 (2006).
15. Leemans, W. *et al.* *Nature Phys.* **2**, 696–699 (2006).
16. Hogan, M. J. *et al.* *Phys. Rev. Lett.* **95**, 054802 (2005).

EVOLUTIONARY BIOLOGY

Caught right-handed

A. Richard Palmer

Are two penises better than one? Not so, implies a study of doubly endowed earwigs. An ancestral behavioural preference for the right penis might have facilitated the loss of the left in species that arose later.

Human males may sometimes wonder about the size of their penis, but they rarely fret about which one to use. Not so for many arthropods, among them fairy shrimp¹, dragonflies² and spiders³, some of which face a delicate choice before each tryst: “Left or right tonight?” Doubly endowed lizards⁴ and snakes⁵ seem ambivalent, although they tend to alternate between the right and left of their two members. Males of two theridiid spider genera (*Tidarren* and *Echinotheridion*) have a particularly radical

solution to their quandary: they voluntarily rip off a palp at random and eat it, leaving only one behind.

Writing in the *Journal of Morphology*⁶, Yoshitaka Kamimura describes his investigations of the private life of the doubly endowed male of the earwig species *Labidura riparia* (Fig. 1a). He shows that this earwig has a strong preference for its right penis: nearly 90% of field-collected and laboratory-reared males hold their intromittent organs in the ‘right-ready’

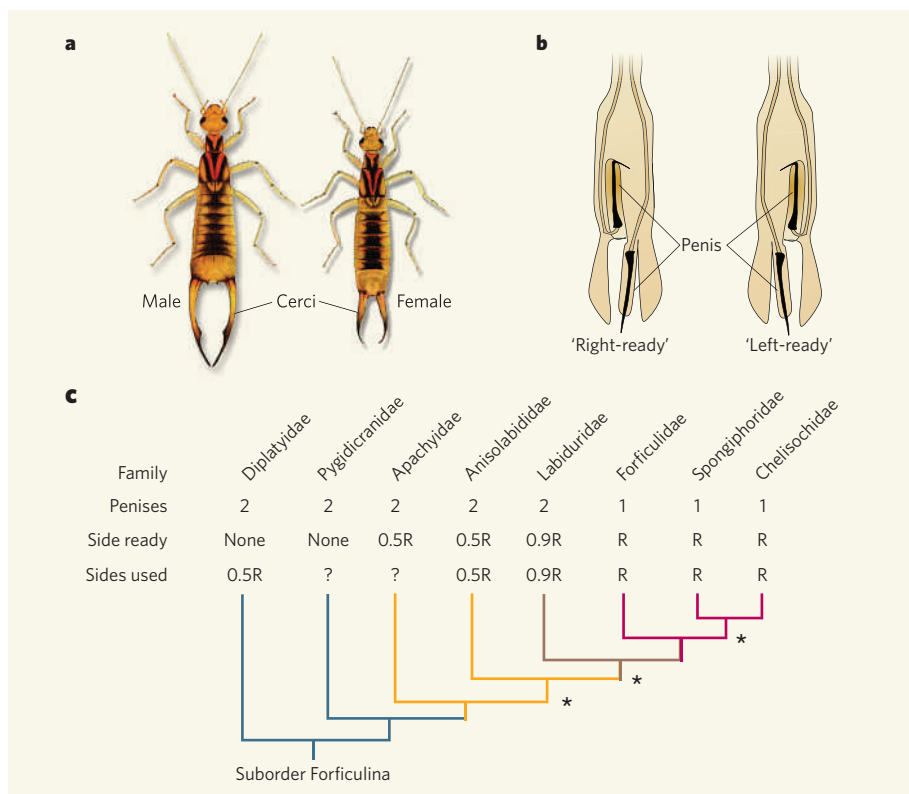


Figure 1 | Earwig penis number, orientation and evolution. **a**, A male and female *Labidura riparia*, the object of Kamimura's studies⁶, showing the pronounced cerci for which earwigs are best known. **b**, The genitalia of the doubly endowed male *L. riparia* showing two different penis orientations while at rest (not mating): 'right ready' (left penis folded anteriorly) and 'left-ready' (right penis folded anteriorly). Nine out of ten *L. riparia* prefer right-ready. (Figure modified from ref. 6.) **c**, Tentative phylogenetic relations among the families of the earwig suborder Forficulina^{7,8}, with asterisks indicating nodes with weak or inconsistent support. The inconsistencies are resolved here following ref. 7. Number of penises, and 'ready' and 'used' side is shown (0.5R = right-side penis ready or used in 50% of cases, and so on; data from ref. 6). The coloured lines represent different combinations of behaviours as indicated by the data above each line.

state (right side extended backwards, ready to mate; Fig. 1b) when not mating, as well as when *in flagrante delicto*. Curiously, the earwig's two penises are morphologically indistinguishable and fully functional. They connect to equivalent testes, and individuals with an injured or experimentally ablated right penis readily revert to using the left one. This right-ready asymmetry is therefore largely — if not entirely — behavioural.

On its own, right-handed penis preference in doubly endowed earwigs would qualify only as amusing natural history. But two other observations noted by Kamimura⁶ increase its significance. The first is that not all earwig taxa have two penises. Some have only one, and it is indeed on the right side. The second is that not all earwigs with two penises have a right-side preference. Some (such as certain species of the families Diplatyidae and Anisolabididae) use the right or left indifferently.

It is the evolutionary relations among these earwigs^{6–8} that elevates a genitalic curiosity to a higher plane (Fig. 1c). First, the earliest two lineages of the earwig suborder Forficulina, the families Diplatyidae and Pygidicranidae, have two penises, both of which point forward in the 'not-ready' position when not mating.

Second, the three families Apachyidae, Anisolabididae and Labiduridae, in which doubly endowed males always hold one penis, either left or right, in the 'ready' position when at rest, are closely related, and lie evolutionarily between the primitive doubly endowed earwigs and the more recently evolved singly endowed earwigs. Third, these later earwigs (the three families Forficulidae, Spongiphoridae and Chelisochidae) that possess only one penis — the right — form a monophyletic group (they are all descendants of a common ancestor) derived from doubly endowed earwigs. Although morphological⁷ and molecular⁸ data yield somewhat different placements, the family Labiduridae, to which the right-leaning *L. riparia* belongs, is closely related to this common ancestor of single-penis earwigs.

Kamimura suggests⁶ an intriguing evolutionary scenario that stems from these phylogenetic relations. Male earwigs evolved from a primitive state with both penises held in the 'not-ready' orientation when not mating, first through a stage where they always held one penis — either the right or left at random — in the 'ready' orientation. The next evolutionary step was males that still possessed two morphologically indistinguishable penises,

but which preferentially held the right in the 'ready' orientation. Finally, the less-preferred (left) penis disappeared altogether, leaving only traces of a closed, non-functional ejaculatory duct. Thus, a purely behavioural asymmetry might have facilitated the evolution of a fully-blown morphological asymmetry.

The current view of the Forficulina phylogeny is in places tentative (Fig. 1c), as is the exact placing of *L. riparia* in its family, so the possibility that right-handed behaviour and left-side penis loss evolved independently cannot be discounted. Nevertheless, the observations seem to support a still controversial and underappreciated mode of evolution, known as the Baldwin effect, that was advanced more than a century ago^{9,10}. This theory of 'organic selection', named after the experimental psychologist James Mark Baldwin, emphasized the impact of behaviour on evolution. Baldwin argued, as had Jean-Baptiste Lamarck nearly a century before¹¹, that new behaviours, learnt or otherwise, can expose an organism to novel conditions of growth to which it may then respond. But unlike Lamarck, Baldwin observed that altered behaviours can also yield altered morphologies through the inherent variation in possible outcomes (the 'plasticity') of an organism's development. If these behaviour-induced morphological differences affect performance positively, natural selection should favour genetic variants that increase or fix the plastic response.

Learnt handedness, coupled with developmental plasticity, provides an attractive mechanism for inducing right-left morphological asymmetries: excess use of one side of the body can induce its overdevelopment, as it does in the upper arm bones of professional tennis players¹². The right-handed *L. riparia* seems to be one of those rare missing links where a right-side behavioural preference has been 'caught' genetically before any morphological differentiation. (As Kamimura shows⁶, *L. riparia*'s right-side preference develops within four days of adult emergence, before any mating experience, so it involves no learning.)

The progression of the story would be deliciously complete if we also knew that individual males of the ambidextrous families Apachyidae and Anisolabididae actually learn to use one side preferentially with increased mating experience. Then the learnt handedness of an individual earwig would clearly have preceded evolutionarily the genetically captured handedness seen in *L. riparia*.

But a puzzling question remains: why a right-side preference? A minor anatomical asymmetry in the female *L. riparia* reproductive tract might favour right-handed males⁶, rather as it does in birds. Most male birds lack an intromittent organ, and mate by pressing together their posterior opening, known as a cloaca, with that of the female. In two species, this 'cloacal kiss' comes most often from the left side^{13,14}, probably because the female's solitary ovary, and the opening to the oviduct, lie to the left¹⁵. But



50 YEARS AGO

The major part of the inaugural address as president of the Institution of Electrical Engineers, delivered by Sir Gordon Radley... was devoted to a consideration of world telecommunication, a subject particularly topical in view of the recent opening to traffic of the first trans-Atlantic telephone cable. The practical realization of inter-continental multi-channel telephone communication by submarine cable is, in Sir Gordon's view, revolutionary in its possibilities, because the traffic capacities of such cables are likely to exceed the existing demand for telegraph and telephone facilities on their routes... A new facility is planned for the trunk telephone network of Great Britain in the form of subscriber-dialling of long-distance calls. To render this practicable, it is necessary to have a nation-wide number scheme...
From *Nature* 8 December 1956.

100 YEARS AGO

To conclude, one may quote some admirable remarks... on the unfortunate result of ignorant European interference with Kafir customs... "In olden days there were regular courts of investigation, consisting of a dozen old women of the kraal. All the girls were medically examined by these women before and after large dances; and thus certain forms of vice were impossible as they would be so speedily detected. Nowadays the young women will not submit to such examination... [A]ncient restraints have been removed, and no new ones have been substituted by white men. The result is disastrous... The case of 'mixed bathing' of the children is another example of a somewhat similar thing. According to Western conceptions of morality this practice is indelicate and liable to lead to immorality. So missionaries advised natives to abandon it. The natives now declare that the abandonment of this custom has led to an increase of immorality, and say that it introduces new vices amongst the people."
From *Nature* 6 December 1906.

Kamimura shows that right-handed and left-handed *L. riparia* have equal mating success⁶.

Alternatively, right-side penis preference might somehow be connected to the unknown factor that favours a stronger curvature of the right cerci in some related wingless earwigs, for example *Anisolabis* and *Euborellia*¹⁶. But the cerci in *L. riparia* are symmetrical (Fig. 1a), so the puzzle remains.

Clearly, the earwig penis system warrants more study. It could become a textbook example of how a possibly learnt lateralized behaviour bred a lateralized morphology evolutionarily. It already qualifies as a fine example of a phenotype-precedes-genotype mode of evolution because the right-ready and left-ready penis variants, which are equally common in evolutionary intermediates (Fig. 1c), and therefore probably not heritable¹⁷, clearly existed before the genetically captured right-ready phenotype seen in *L. riparia*. Who would have ever thought you could learn so much from earwig penises? ■

A. Richard Palmer is in the Department of Biological Sciences, University of Alberta, Edmonton, Alberta, T6G 2E9 Canada.

e-mail: rich.palmer@ualberta.ca

1. Rogers, D. C. & Fugate, M. *West. N. Am. Nat.* **61**, 11–18 (2001).
2. Bechly, G., Brauckmann, C., Zessin, W. & Groning, E. *J. Zool. Syst. Evol. Res.* **39**, 209–226 (2001).
3. Knoflach, B. & van Harten, A. *J. Nat. Hist.* **40**, 1483–1616 (2006).
4. Tokarz, R. R. & Slowinski, J. B. *Anim. Behav.* **40**, 374–379 (1990).
5. Shine, R., Olsson, M. M., LeMaster, M. P., Moore, I. T. & Mason, R. T. *Behav. Ecol.* **11**, 411–415 (2000).
6. Kamimura, Y. *J. Morphol.* **267**, 1381–1389 (2006).
7. Haas, F. & Kukulova-Peck, J. *Eur. J. Entomol.* **98**, 445–509 (2001).
8. Jarvis, K. J., Haas, F. & Whiting, M. F. *Syst. Entomol.* **30**, 442–453 (2005).
9. Baldwin, J. M. *Am. Nat.* **30**, 441–451 (1896).
10. Baldwin, J. M. *Am. Nat.* **30**, 536–553 (1896).
11. Lamarck, J. B. *Philosophie Zoologique* (Dentu, Paris, 1809).
12. Trinkaus, E., Churchill, S. E. & Ruff, C. B. *Am. J. Phys. Anthropol.* **93**, 1–34 (1994).
13. Nyland, K. B., Lombardo, M. P. & Thorpe, P. A. *Wilson Bull.* **115**, 470–473 (2003).
14. Petersen, A. D., Lombardo, M. P. & Power, H. W. *Anim. Behav.* **62**, 739–741 (2001).
15. King, A. S. in *Form and Function in Birds* Vol. 2 (eds King, A. S. & McLelland, J.) 107–148 (Academic, New York, 1981).
16. White, R. E., Borror, D. J. & Peterson, R. T. *A Field Guide to Insects: America North of Mexico* (Houghton Mifflin, Boston, 1998).
17. Palmer, A. R. *Science* **306**, 828–833 (2004).

CHEMICAL BIOLOGY

Renewing embryonic stem cells

Reka R. Letso and Brent R. Stockwell

Embryonic stem cells have great potential in medicine, but the current methods used to grow them prevent their therapeutic use. A dual-action compound has been discovered that may help solve this problem.

Last year, the International Stem Cell Forum stated that a reliable method for growing stem cells that does not depend on animal-derived products is a requirement for future work¹. Reporting in *Proceedings of the National Academy of Sciences*, Chen *et al.*² describe their progress towards this goal. Using high-throughput screening, they have discovered a chemical that allows mouse embryonic stem (ES) cells to perpetuate themselves. This compound is a valuable tool for studying ES cell self-renewal, and may bring us closer to developing these cells for therapeutic purposes.

Great hopes have been raised that ES cells will one day be used to replace damaged cells, and to provide therapies beyond the reach of conventional drugs. However, a serious obstacle to their use is our lack of insight into the mechanisms that regulate a stem cell's behaviour — more specifically, whether it undergoes self-renewal or differentiates to become a more specialized type of cell. Furthermore, most human stem-cell lines, including all human ES cell lines approved for study with US federal funds, were grown using animal products, so the potential for cross-species contamination is too high for these lines to be developed for

therapeutic purposes³. A synthetic compound that promotes ES cell self-renewal could help to address both of these issues.

Intriguingly, Chen and colleagues' compound — named SC1 — hits two targets in a protein network. Both of these targets are necessary to promote ES cell self-renewal. The authors discovered the compound by high-throughput screening of a chemical library⁴ they had designed to target kinase enzymes; kinases pass on signals inside the cell. The authors tested their chemicals in mouse ES cells that were engineered to produce green fluorescent protein (GFP) only when they were perpetuating themselves. For the screen, the cells were grown under conditions that should cause them to stop self-renewing and to differentiate, a process that causes the levels of GFP to drop and the fluorescence to disappear. Chen *et al.* looked for compounds that had the ability to maintain fluorescence under these conditions.

Surprisingly, although the library was designed to target kinase enzymes, only one of the two targets of SC1 is a kinase. This illustrates an interesting notion in the design of chemical libraries: certain chemical groups

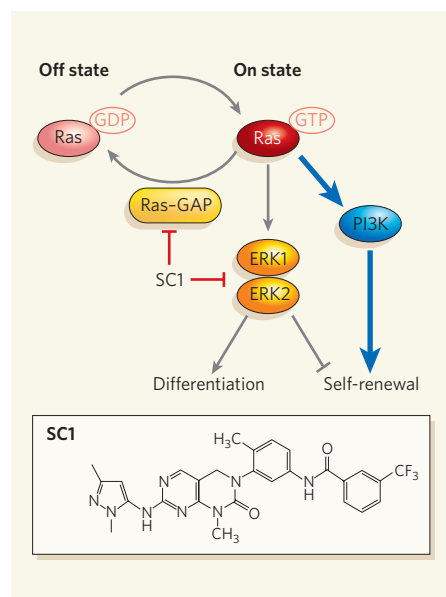


Figure 1 | Chemically induced stem-cell self-renewal. Chen *et al.*² have discovered a synthetic compound, SC1, that interferes with signalling in mouse embryonic stem cells. In a cell, the Ras protein is activated by binding guanosine triphosphate (GTP). This in turn activates the enzyme PI3K, which promotes stem-cell self-renewal. Activated Ras also switches on the enzymes ERK1 and ERK2, which promote differentiation of the cell. Ras can be deactivated by the enzyme Ras-GAP, which converts GTP to guanosine diphosphate (GDP). SC1 inhibits Ras-GAP, so that Ras remains activated, enhancing stem-cell renewal via the PI3K pathway. SC1 also inhibits ERK1 and ERK2, thus blocking stem-cell differentiation.

bind well to a variety of cellular proteins, even though the structures of those proteins may be very different. In addition, SC1 highlights the value of using small molecules in screens that probe signalling pathways as an alternative to genetic methods, such as RNA interference or complementary DNA expression. As the authors explain², the “advantage (and complexity) in the use of small molecules...is that compounds can modulate more than one target to achieve a desired biological effect”. Moreover, using current technology for protein identification, one can identify and dissect these multiple relevant targets, as the authors did here.

Chen *et al.* found that the two proteins targeted by SC1 are in the Ras signalling network (Fig. 1). Ras proteins are principal players in a complex network that regulates many cellular functions, including growth and mobility. These proteins stimulate growth when activated, that is, when they are bound to the molecule guanosine triphosphate (GTP). One of the targets of SC1 — Ras-GAP — is a protein that modulates Ras activity by stimulating the Ras protein to convert its GTP into guanosine diphosphate, turning off Ras signalling. So, inhibition of Ras-GAP by SC1 increases Ras signalling; Ras then activates its signalling partners farther along the network — including the kinase enzyme

PI3K, which has been implicated in stem-cell self-renewal⁵. Think of the old aphorism “the enemy of my enemy is my friend”; SC1 inhibits the inhibitor of Ras, thereby activating Ras and PI3K signalling.

However, activation of the entire Ras network is unlikely to result in stem-cell self-renewal, as two aspects of Ras signalling oppose each other in regulating perpetuation. The PI3K arm of the Ras network promotes self-renewal, whereas another arm — containing the kinases ERK1 and ERK2 — inhibits self-renewal in mouse ES cells. By inhibiting ERK1 and ERK2 signalling, SC1 funnels the activated Ras signal towards the PI3K arm. This is a sophisticated effect for a small molecule, and it illustrates the power of this screening approach to identify unexpected ways of intervening in biological systems.

The authors have yet to test SC1 in human embryonic or adult stem cells, and it is possible that differences between signalling pathways in mice and humans may prevent SC1 from being active in human cell lines. For example, inhibition of ERK1 and ERK2 signalling in mouse ES cells promotes self-renewal, but the opposite is

true for human ES cells. However, even if SC1 is inactive in human cell lines, the discovery of a compound that can promote self-renewal in mouse cells strongly suggests that there may be a parallel pathway to exploit in their human counterparts. This would simplify the method of growing human ES cells, provide information about the mechanisms controlling self-renewal, and offer a contaminant-free method of growing these cells for therapeutic purposes.

Reka R. Letso and Brent R. Stockwell are in the Department of Biological Sciences and the Department of Chemistry, Sherman Fairchild Center for the Life Sciences 614, Columbia University, 1212 Amsterdam Avenue, MC 2406, New York, New York 10027, USA.
e-mail: stockwell@biology.columbia.edu

1. Moore, H. *Nature Biotechnol.* **24**, 160–161 (2006).
2. Chen, S. *et al. Proc. Natl Acad. Sci. USA* **103**, 17266–17271 (2006).
3. Martin, M. J., Muotri, A., Gage, F. & Varki, A. *Nature Med.* **11**, 228–232 (2005).
4. Ding, S., Gray, N. S., Wu, X., Ding, Q. & Schultz, P. G. *J. Am. Chem. Soc.* **124**, 1594–1596 (2002).
5. Armstrong, L. *et al. Hum. Mol. Genet.* **15**, 1894–1913 (2006).

CONDENSED-MATTER PHYSICS

Defects and perfect flows

Henry R. Glyde

The discovery that parts of a solid helium crystal could flow through other parts without friction ignited physicists' interest. Independent experiments confirm this unusual superflow, but its origin remains mysterious.

In 2004, Kim and Chan reported the spectacular, and controversial, observation of superfluidity — flow without resistance from frictional forces — in crystalline helium^{1,2}. This remarkable finding has now been confirmed^{3–6}. But the latest experiments indicate that, rather than being an intrinsic property of a perfect quantum solid, superflows owe their existence to macroscopic defects or extended disorder in the structure of solid helium.

Essentially, Kim and Chan observed that a small fraction (around 1%) of the solid ⁴He mass decoupled from the rest of the solid below a critical temperature, T_C , of around 0.2 K. This component, denoted the superfluid fraction, ρ_s , could flow (or remain at rest) without friction inside the solid. The piquancy of the discovery lay in its extending the concept of superfluidity to all three phases of matter: gases, liquids and solids. Before 2004, superflows had been observed only in fluids. These systems include paired electrons in solids, the root cause of superconductivity (discovered in 1911); ⁴He atoms in the liquid state (1938); and paired ³He atoms in liquid ³He (1972). The possibility that superflow could occur in a lattice whose atoms have long-range positional

order — a crystal — seemed remote.

In three new measurements^{3–5}, the rotational moment of inertia of solid ⁴He is determined in a rotating apparatus known as a torsional oscillator. The decoupling of the superfluid component appears as a reduced moment below T_C , at which point a fraction of the solid ceases to oscillate with the rest. Writing in *Physical Review Letters*, Kim and Chan³ confirm their earlier observation¹ that the magnitude of ρ_s varies from sample to sample of solid ⁴He (Fig. 1). In the same journal, Rittner and Reppe⁴ find that ρ_s can be substantially reduced, even to zero, if the solid sample is annealed; that is, if it is warmed and then re-cooled to below T_C . Shirahama and co-workers⁵ have reported similarly that the supersolid fraction can be halved — but not eliminated entirely — by annealing. The implication of all these experiments is that superfluid decoupling depends on imperfections in the solid helium, as these would vary between samples and would be modified by annealing.

Soon after Kim and Chan's initial discovery^{1,2}, efforts were made to observe superflow directly by pressing solid helium against a barrier containing small pores⁷. The superfluid

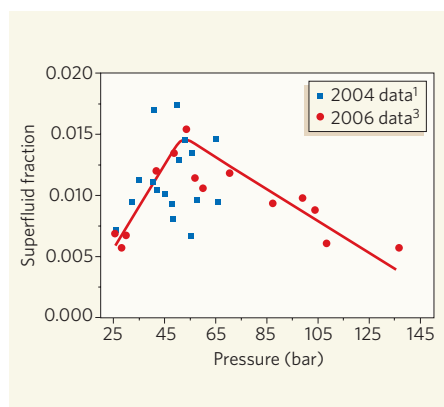


Figure 1 | Supersolid component. The superfluid fraction in crystalline ^4He in the low-temperature limit, as a function of pressure, from Kim and Chan's data^{1,3}. Crystals grown at constant pressure and temperature (2006 data)³ show somewhat less variation in the supersolid fraction than do those grown at constant volume (2004 data)¹. The fraction also does not decrease sharply with increasing pressure, as would be expected if the phenomenon were the result of quantum-mechanical exchange processes in a perfect crystal. The solid line is a guide to the eye.

component should have flowed immediately through the fine pores, but no superflow was observed. This year, however, Sasaki and colleagues⁶ have observed bulk superflow. Again, it is not seen in all samples, but only in those that have a large, observable boundary between two grains (regions of distinct crystallographic orientations) extending across the sample. Superflow seems to occur along, or close to, these grain boundaries.

To understand the significance of these results^{3–6}, one must first know that there are two prevalent pictures of superfluidity. The first of these focuses on the quantum-mechanical exchange, or tunnelling, of atoms between lattice sites. At low temperature, the de Broglie wavelength of ^4He atoms — a quantum-mechanical measure of a particle's extent in space — is long and covers many sites. A long wavelength enables long-range exchange of atoms between lattice sites. If the exchanges extend right across the sample, superfluidity occurs.

The second picture is based on the phenomenon of Bose–Einstein condensation. Any number of the particles known as bosons, which possess integer spin, may occupy a single quantum state. If, at low temperature, a macroscopic fraction of bosons condenses into a single state, a long-range coherence is brought to the system that makes superfluidity possible. In liquid ^4He , the classic superfluid, Bose–Einstein condensation is known to occur below the critical temperature for superfluidity in this system, 2.17 K. As the temperature approaches absolute zero — where ρ_s is 100% — the condensed fraction is around 7%.

Early theoretical discussions^{8–10} of possible superflow in solid ^4He involved quantum tunnelling through ground-state vacancies (these

are lattice sites that are vacant at absolute zero), as well as Bose–Einstein condensation and quantum exchanges within the lattice. The predicted¹⁰ superfluid fraction within the lattice was of the order 0.01%. Unfortunately, however, ground-state vacancies have not been observed in solid ^4He . Thermally activated vacancies are found — but at temperatures above T_C that are irrelevant to the establishment of superfluidity. Recent calculations¹¹ also find that individual vacancies are not stable in the ground state of crystalline ^4He . The vacancies instead coalesce or migrate to a surface; this is the mechanism by which thermal vacancies leave a classical crystal when it is cooled.

Calculations have also shown that there is no long-range coherence, and so no Bose–Einstein condensation, in a perfect crystal of ^4He , essentially because condensation requires double occupancy of a lattice site^{12,13}. In real crystals, the condensate fraction is observed in neutron-scattering experiments¹⁴ to be $0.2 \pm 0.6\%$ — that is, compatible with zero — at a temperature of 0.08 K. Similarly, long-range quantum exchanges within crystalline ^4He are too infrequent to explain the high ρ_s observed¹⁵. As seen in solid ^3He , quantum exchange rates decrease dramatically as a solid is compressed under pressure. If the superfluid fraction arises from exchanges of particles between lattice sites, ρ_s should decrease by orders of magnitude under increasing pressure. But Kim and Chan³ find ρ_s to be largely independent of pressure (Fig. 1). Superflow in crystalline ^4He thus does not seem to be a phenomenon of the perfect bulk solid^{11–13,15}, or to involve individual point defects (vacancies) that are in equilibrium at absolute zero.

Given the recently discovered dependence

of ρ_s on sample annealing^{3–5}, and the correlation of superflow with grain boundaries⁶, supersolidity seems to hinge on macroscopic, long-range defects such as grain boundaries or amorphous channels in the helium crystal. But at the same time, a superfluid density of the same magnitude is observed¹ in helium confined in a nanoporous medium such as Vycor. It is difficult to imagine that extended defects could be the same in helium thus confined and in bulk helium. The situation is far from clear: revealing the secrets of this latest superfluid is very much a work in progress. ■ Henry R. Glyde is in the Department of Physics and Astronomy, University of Delaware, Newark, Delaware 19716, USA.

e-mail: glyde@udel.edu

- Kim, E. & Chan, M. H. W. *Nature* **427**, 225–227 (2004).
- Kim, E. & Chan, M. H. W. *Science* **305**, 1941–1944 (2004).
- Kim, E. & Chan, M. H. W. *Phys. Rev. Lett.* **97**, 115302 (2006).
- Rittner, A. S. C. & Reppey, J. D. *Phys. Rev. Lett.* **97**, 165301 (2006).
- Shirahama, K., Kondo, M., Takada, S. & Shibayama, Y. *Am. Phys. Soc. March Meet. Abstr.* G41.00007 (2006).
- Sasaki, S., Ishiguro, R., Caupin, F., Maris, H. J. & Balibar, S. *Science* **313**, 1098–1100 (2006).
- Day, J. & Beamish, J. *Phys. Rev. Lett.* **95**, 105304 (2006).
- Andreev, A. F. & Lifshitz, I. M. *Sov. Phys. JETP* **29**, 1107–1113 (1969).
- Chester, G. V. *Phys. Rev. A* **2**, 256–258 (1970).
- Leggett, A. J. *Phys. Rev. Lett.* **25**, 1543–1546 (1970).
- Boninsegni, M. *Phys. Rev. Lett.* **97**, 080401 (2006).
- Boninsegni, M., Prokof'ev, N. & Svistunov, B. *Phys. Rev. Lett.* **96**, 105301 (2006).
- Clark, B. K. & Ceperley, D. M. *Phys. Rev. Lett.* **96**, 105302 (2006).
- Diallo, S. O., Pearce, J. V., Taylor, J. W., Kirichek, O. & Glyde, H. R. *Quantum Fluids and Solids Symp. Kyoto, August 2006* (in the press).
- Ceperley, D. M. & Bernu, B. *Phys. Rev. Lett.* **93**, 155303 (2004).

OCEANOGRAPHY

Plankton in a warmer world

Scott C. Doney

Satellite data show that phytoplankton biomass and growth generally decline as the oceans' surface waters warm up. Is this trend, seen over the past decade, a harbinger of the future for marine ecosystems?

Oranges in Florida, wildfires in Indonesia, plankton in the North Pacific — what links these seemingly disparate items is that they are all affected by year-to-year fluctuations in global-scale climate. On page 752 of this issue, Behrenfeld *et al.*¹ describe how such fluctuations, especially in temperature, are connected to the productivity of phytoplankton in the world's oceans. Their analyses are based on nearly a decade of satellite data, and for much of the oceans they find that recent warmer surface temperatures correspond to lower oceanic biomass and productivity. Behrenfeld *et al.* argue that these patterns arise because

climate-induced changes in ocean circulation reduce the supply of nutrients needed for photosynthesis.

Small photosynthetic phytoplankton grow in the well-illuminated upper ocean, forming the base of the marine food web and supporting the fish stocks we harvest. They also form the basis of the biogeochemical cycling of carbon and many other elements in the sea. Phytoplankton growth depends on temperature and the availability of light and nutrients, including nitrogen, phosphorus, silicon and iron. Most of this nutrient supply to the surface ocean comes from the mixing and upwelling of cold, nutrient-rich

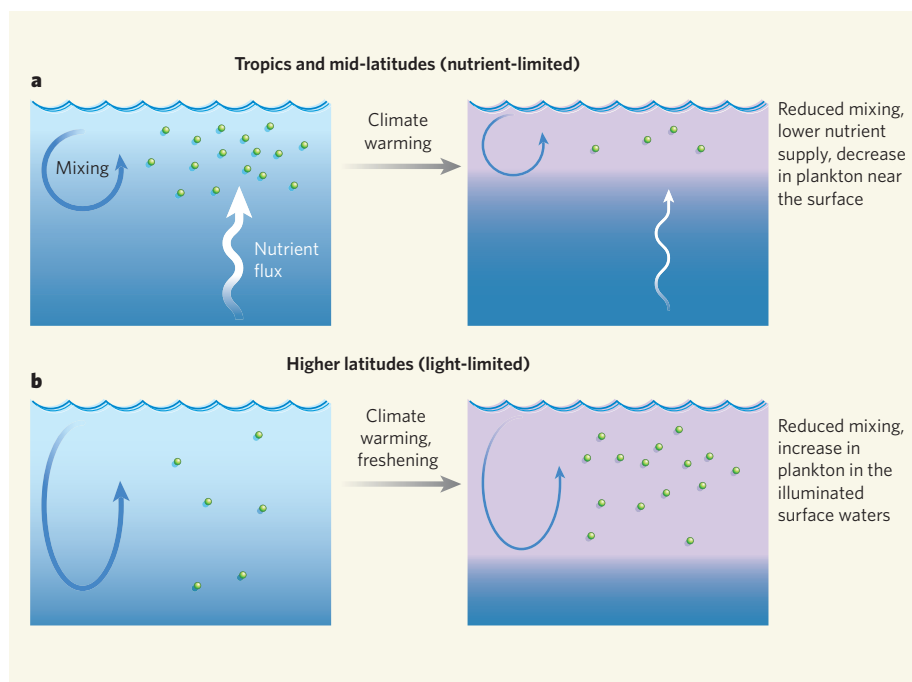


Figure 1 | Predicted phytoplankton response to increased temperature in ocean surface waters¹.

a, In the tropics and at mid-latitudes, phytoplankton are typically nutrient-limited, and satellite data tie reduced biological productivity to upper-ocean warming and reduced nutrient supply. **b**, At higher latitudes, the opposite biological response to future warming, and extra freshwater input, may occur. In these regions, phytoplankton are often light-limited; reduced mixing would keep plankton close to the surface where light levels are higher.

water from below, with an additional source of iron from mineral dust swept off the continental deserts. Phytoplankton biomass can vary by a factor of 100 in surface waters; its geographical distribution is determined largely by ocean circulation and upwelling, with the highest levels being found along the Equator, in temperate and polar latitudes, and along the western boundaries of continents.

Although the broad spatial patterns of phytoplankton biomass and productivity are well documented², large-scale temporal variations have only recently become quantifiable with the advent of satellite ocean-colour sensors³. The ocean is vast, and the limited number of research ships move at about the speed of a bicycle, too slow to map the ocean routinely on ocean-basin to global scales. By contrast, a satellite can observe the entire globe, at least the cloud-free areas, in a few days. Phytoplankton biomass and growth rates can be estimated remotely from space because chlorophyll, the main photosynthetic pigment in phytoplankton, absorbs blue and red sunlight more readily than green sunlight. Ocean-colour sensors measure, by wavelength band, the small fraction of sunlight scattered back to space from below the surface. The resulting surface-chlorophyll data can be combined with empirical relationships to estimate phytoplankton growth rates or net primary production⁴.

Not that this procedure is straightforward: other constituents of sea water absorb light; many photons reaching the satellite sensor come from atmospheric aerosols or reflection at the water surface; and optical detectors on

satellites degrade with time. But with careful calibration, high-quality, long-term records of ocean-colour data can be constructed for detecting climate-driven trends. The best such record at present is from GeoEye and from NASA's Sea-viewing Wide Field-of-view Sensor (SeaWiFS)³, launched in 1997. In the SeaWiFS time series, global chlorophyll and productivity increase sharply during 1997–98 and then decline gradually to 2005.

Behrenfeld *et al.*¹ show that these trends closely follow changes in climate. The phytoplankton increase in 1997–98 matches a negative (cold) phase of the El Niño–Southern Oscillation (ENSO), and the subsequent slow drop occurred as the planet moved into an extended warm ENSO period. ENSO is a dominant mode of interannual climate variability, and involves large-scale reorganizations of atmosphere and ocean circulation that originate in the equatorial Pacific and extend across the globe. Decreases in productivity are equally well related to increases in sea surface temperatures and vertical temperature gradients in the upper ocean.

The climate–plankton link is found primarily in the tropics and mid-latitudes, where there is limited vertical mixing because the water column is stabilized by thermal stratification (that is, when light, warm waters overlie dense, cold waters). In these areas, the typically low levels of surface nutrients limit phytoplankton growth. Climate warming further inhibits mixing, reducing the upward nutrient supply and lowering productivity (Fig. 1a). At higher latitudes, phytoplankton

are often light-limited because intense vertical mixing carries them hundreds of metres down into darkness where sunlight does not penetrate. In these regions, future warming and a greater influx of fresh water, mostly from increased precipitation and melting sea-ice, will contribute to reduced mixing that may actually increase productivity⁵ (Fig. 1b).

The recent observed global surface warming (about 0.2 °C per decade) is expected to accelerate in coming decades as we continue to release excess carbon dioxide into the atmosphere. In fact, the planet may soon be warmer than at any time in the past million years⁶. Extrapolating the satellite observations into the future suggests that marine biological productivity in the tropics and mid-latitudes will decline substantially, in agreement with climate-model simulations^{7–9}. In those simulations, the geographical boundaries that separate specific marine ecosystems (the ocean equivalents of forests, grasslands and so on) migrate towards the poles, and productivity increases at high latitudes because of surface warming, enhanced freshwater input and reduced deep mixing. Ecosystem dynamics are complex and nonlinear, however, and unexpected phenomena may arise as we push the planet into this unknown climate state. For example, oceanic fixation of atmospheric nitrogen into biologically available forms is concentrated in warm, nutrient-poor surface waters; under more stratified conditions, fixation might increase and enhance overall productivity^{8,10}.

Detecting the impact of climate change requires increased and more sophisticated monitoring of ocean biology and environmental change, both from space and with instruments in the water. Ideally, we need measures not only of such parameters as chlorophyll concentration and productivity, but also of plankton taxonomy (what is there?) and physiology (how healthy are they?). New remote-sensing technologies may help meet these challenges¹¹.

Scott C. Doney is in the Marine Chemistry and Geochemistry Department, Woods Hole Oceanographic Institution, Woods Hole, Massachusetts 02543, USA.
e-mail: sdoney@whoi.edu

- Behrenfeld, M. J. *et al.* *Nature* **444**, 752–755 (2006).
- Longhurst, A. *Ecological Geography of the Sea* 2nd edn (Academic, New York, 2006).
- McClain, C. R., Feldman, G. C. & Hooker, S. B. *Deep-Sea Res.* **51**, 5–42 (2004).
- Carr, M. E. *et al.* *Deep-Sea Res.* **53**, 741–770 (2006).
- Polovina, J. J., Mitchum, G. T. & Evans, G. T. *Deep-Sea Res.* **42**, 1701–1716 (1995).
- Hansen, J. *et al.* *Proc. Natl Acad. Sci. USA* **103**, 14288–14293 (2006).
- Bopp, L. *et al.* *Glob. Biogeochem. Cycles* **15**, 81–99 (2001).
- Boyd, P. W. & Doney, S. C. *Geophys. Res. Lett.* **29**, 1806 (2002).
- Sarmiento, J. L. *et al.* *Glob. Biogeochem. Cycles* **18**, GB3003; doi:10.1029/2003GB002134 (2004).
- Karl, D. *et al.* *Nature* **388**, 533–538 (1997).
- NASA OBB Advanced Planning Team *Plan for Earth's Living Ocean: The Unseen World to Set NASA's Ocean Biology and Biogeochemistry Agenda* (NASA, in the press).

ION CHANNELS

A paddle in oil

Anthony G. Lee

How do voltage-gated ion channels in cell membranes open? The latest work suggests that the process depends on having the correct lipid molecules in the membrane, with phosphate groups being mandatory.

Electrical signalling in the nervous system involves voltage-gated ion channels, proteins that sit in the outer membrane of nerve cells. This membrane, like all membranes, consists of lipid molecules that are organized in the form of a bilayer — the charged parts of the molecules, the lipid headgroups, lie on the outside of the bilayer in contact with water, with the fatty acyl chains of the lipid molecules occupying the centre and forming the oily core. Embedded in this lipid bilayer are the voltage-gated ion channels. On page 775 of this issue, Schmidt, Jiang and MacKinnon¹ show that having the correct lipids in the membrane is essential if the ion channels are to open when the voltage changes across the membrane*.

The voltage change that opens an ion channel — about 50 mV — seems quite small, but a voltage change of 50 mV across a membrane 50 Å thick corresponds to a change of about $100,000 \text{ V cm}^{-1}$, easily enough to cause a change in the structure of an ion channel. Our views of what exactly this structural change might be altered dramatically when MacKinnon's group determined the X-ray structure of a voltage-gated ion channel^{2,3}. The part of the ion channel that senses the change in voltage across the membrane is highly positively charged and is loosely attached to the outer face of the channel, exposed to the lipid bilayer. Because of its shape, MacKinnon referred to this part of the channel as a paddle (Fig. 1). The idea is that a change in voltage across the membrane results in movement of the paddle, which in turn leads to opening of the central pore of the channel, allowing ions to flow through the pore. It is currently a matter of debate whether the paddle moves all the way across the lipid bilayer when the voltage changes or whether its motion is more restricted⁴.

It seems likely that a bilayer of the correct lipid composition is essential for proper function of the paddle, because the paddle is exposed to the bilayer and has to move in it. The idea that the lipids are important has now been investigated by MacKinnon and colleagues¹ in a very direct way. They used the traditional approach of forming lipid-bilayer membranes across a small hole in a plastic partition, and then inserting a voltage-gated

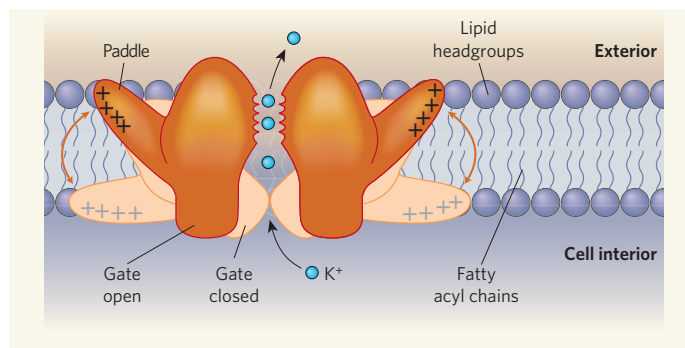


Figure 1 | Lipids and voltage-gated potassium channels. Voltage-gated potassium channels contain positively charged 'paddles', loosely associated with the surface of the channel, that move within the lipid bilayer to open the channel. After ruling out some explanations for what bilayer composition is required for correct paddle operation, MacKinnon and colleagues¹ conclude that negatively charged phosphate groups are an essential component.

ion channel into the membrane by fusing vesicles containing the purified channel with the membrane. This allowed them to measure the current flowing through the ion channel as a function of the lipid composition of the membrane. For their experiments, they used the voltage-gated potassium channel from the bacterium *Aeropyrum pernix* because this can be expressed in high quantities in *Escherichia coli* and then purified.

The lipid composition chosen by MacKinnon and colleagues¹ for their initial experiments was 70% phosphatidylethanolamine and 30% phosphatidylglycerol. This is the lipid composition of *E. coli* cell membranes, and the ion channel behaved perfectly in a bilayer of these lipids.

But what happens when the lipid composition is changed? The authors looked at the importance of phosphatidylethanolamine. Although this is the most abundant lipid in the *E. coli* membrane, it is rather unusual in that it does not, on its own, form bilayers at normal temperatures. Rather, because the overall shape of the molecule is conical, it tends to pack in three dimensions to give curved, non-bilayer structures. However, the presence in the membrane of a cylindrically shaped lipid molecule such as phosphatidylglycerol forces the phosphatidylethanolamine to adopt a bilayer structure, with the latter molecules being said to exist in the bilayer in a state of 'curvature frustration'. There has been much speculation as to the possible significance of this frustrated state⁵. However, the authors found that channel function was unaltered when phosphatidylethanolamine was substituted with the cylindrically shaped, bilayer-favouring

lipid phosphatidylcholine. So it seems that curvature frustration is not important for ion-channel function.

MacKinnon and colleagues then turned to the effect of phosphatidylglycerol, a lipid molecule with a negatively charged (anionic) headgroup. Anionic lipids would be expected to bind strongly to the positively charged paddle; it has been shown⁶, for example, that negatively charged lipids bind with high affinity to a positively charged patch on the mechanosensitive channel MscL. But the voltage-gated potassium channel was functional in the

absence of phosphatidylglycerol, so any interaction between anionic lipids and the paddle cannot be essential for channel function.

Rather, it turns out that what is essential for function is the presence of a negatively charged phosphate group in the lipid molecule. The channel was totally inactive in non-natural, positively or negatively charged lipid molecules lacking a phosphate group, or in an uncharged, sugar-based lipid. It is intriguing that the membrane lipids of *A. pernix*, from which the channel originally came, are, unusually, all anionic, phosphate-containing lipids⁷.

Why is the phosphate group vital? MacKinnon and colleagues suggest that negatively charged phosphate groups in the lipid molecules may act as counterions for the positively charged arginine amino-acid residues in the paddle, stabilizing the charged paddle in the lipid bilayer. If this idea is correct, it would provide another example of the importance for membrane-protein function of the local interactions between a membrane protein and its surrounding lipid molecules. ■

Anthony G. Lee is in the School of Biological Sciences, University of Southampton, Southampton SO16 7PX, UK.
e-mail: agl@soton.ac.uk

- Schmidt, D., Jiang, Q.-X. & MacKinnon, R. *Nature* **444**, 775–779 (2006).
- Jiang, Y. et al. *Nature* **423**, 33–41 (2003).
- Long, S. B., Campbell, E. B. & MacKinnon, R. *Science* **309**, 897–903 (2005).
- Tombola, F., Pathak, M. M. & Isacoff, E. Y. *Neuron* **48**, 719–725 (2005).
- Lee, A. G. *Biochim. Biophys. Acta* **1666**, 62–87 (2004).
- Powl, A. M., East, J. M. & Lee, A. G. *Biochemistry* **44**, 5873–5883 (2005).
- Morii, H. et al. *Biochim. Biophys. Acta* **1436**, 426–436 (1999).

*This article and the paper concerned¹ were published online on 29 November 2006.

CHEMISTRY

Metals line up for DNA

Jens Müller

The versatile DNA molecule has found many applications beyond biology. In its latest role, it serves as a self-assembling scaffold to arrange different metal ions in a row, like pearls on a string.

The days when DNA was the exclusive purview of molecular biologists are over. Since the discovery of DNA's double-helix structure in 1953, numerous disciplines have embraced this biomolecule — from medicine to materials science, by way of classical chemistry and biotechnology. Reporting in *Nature Nanotechnology*, Tanaka *et al.*¹ describe an application of DNA that clearly stems from inorganic chemistry: they have modified DNA so that it serves as a scaffold for one-dimensional arrays of metal ions. Each array is a combination of copper and mercury ions, arranged in an order defined by the DNA sequence. Such precise control over the assembly of these arrays would be necessary for their potential applications as nanomagnets², as self-assembling molecular wires or as catalysts in chemical reactions.

Single strands of DNA comprise a sugar-phosphate chain decorated with organic bases. A natural DNA double helix consists of two single strands, with the sugar-phosphate backbones on the outside and the bases on the inside (Fig. 1a). Each base forms hydrogen bonds with just one other kind of base — a complementary base — on the opposite strand. It is the complementarity of these bases that drives the reliable self-assembly process of double-helix formation.

Tanaka *et al.*¹ replaced the natural bases of DNA with artificial ones. Each of these substitute bases has a high affinity for a particular metal ion — either a copper ion (Cu^{2+}) or a mercury ion (Hg^{2+}). The modified DNA can form a double helix only if the opposing bases have a preference for the same metal ion and if both bases bind to such an ion. Given an appropriate base sequence, a duplex can form that contains one or more metal ions along its central axis (Fig. 1b).

The incorporation of metal ions into an artificial DNA duplex has previously been reported, but with only one kind of metal at a time, typically in systems with a few modified base pairs interspersed between longer rows of natural ones^{3–8}. What is remarkable about Tanaka and colleagues' work¹ is the formation of double helices containing long stretches of two different metal ions. Not only did the authors selectively incorporate more than one kind of metal ion into artificial DNA, but they also used more than one kind of artificial base, in various sequences. This is a tremendous advance towards the goal of making metal-containing, self-assembling nanomaterials with desirable properties that can be tuned on

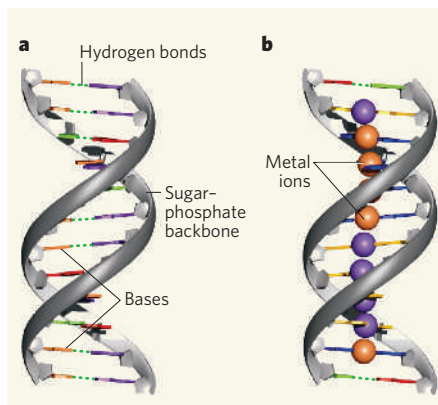


Figure 1 | Artificial DNA that incorporates metal ions. **a**, Single strands of DNA consist of a sugar-phosphate backbone decorated with organic bases. A double helix forms when the bases of one strand form hydrogen bonds to complementary bases on another strand. **b**, Tanaka *et al.*¹ have replaced bases in natural DNA with artificial ones that bind specifically to copper ions (Cu^{2+} , purple spheres) or mercury ions (Hg^{2+} , orange spheres). A duplex forms only when opposing bases bind a metal ion between them. The sequence of metal ions is determined by the sequence of artificial bases.

the basis of the order of the metal ions. The beauty of Tanaka and colleagues' work is that it exploits the self-assembly of DNA, where the two strands come together, a process that has been optimized by evolution over several billion years. Furthermore, automated DNA synthesis permits quick and easy access to any desired sequence of artificial DNA, provided that a certain length of molecule is not exceeded.

So how could these artificial DNA structures be used? One possible application is in organic chemistry, as catalysts for enantioselective reactions⁹ — that is, reactions that proceed with precise control of the three-dimensional structure of the products. Such catalysts are essential for the synthesis of biologically active compounds, but tailoring the interactions between a catalyst and a reactant to achieve this precise control is far from simple. Natural catalysts, such as enzymes, are excellent at directing enantioselective reactions, but from a chemist's perspective their range of reaction conditions is very restricted. Synthetic metal catalysts can be optimized to work under more typical laboratory conditions and on an industrial scale, but they are often less efficient than enzymes. Metal-ion-containing DNA catalysts

represent a biologically inspired approach that could combine the best of both of these catalytic worlds.

It has been found that, in a short artificial DNA duplex containing an array of five consecutive copper ions, the ions interact magnetically with each other²; this helix can be thought of as a self-assembled nanomagnet. Using Tanaka and colleagues' approach¹, specific combinations of metal ions might be incorporated into DNA to fine-tune the magnetic properties of such devices. The electrical properties of DNA are also of interest. Unmodified DNA lacks sufficient electrical conductivity to be used in molecular electronic devices¹⁰. To address this problem, DNA has previously been prepared with metals attached to its exterior, but this modification prevented the DNA from reversibly self-assembling¹¹. Tanaka and colleagues' structures¹ might provide a way forward, as they incorporate metal ions in the centre of a DNA double helix that retains its ability to self-assemble.

But difficulties might be encountered when scaling up the authors' technique to build larger molecules. As DNA duplexes form, intermediate double helices can be produced in which the bases are not perfectly matched. But every metal ion incorporated into the artificial DNA increases the stability of the final aggregate. This effect could stabilize the intermediate helices so much that they no longer rearrange to form a perfect duplex, so introducing errors during DNA self-assembly.

Perhaps more importantly, automated DNA synthesis can only prepare sequences of up to about 100 bases. To build larger structures, biological techniques must be developed that either splice together short artificial DNA strands or allow the synthesis of longer strands. Such methods exist for natural DNA, but these would have to be modified to tolerate artificial bases and metal ions. Finally, more work is needed to determine the electrical properties of metal-ion-containing DNA — is it a conductor or a semiconductor? Future research should focus on this question, and investigate the factors that influence conductivity. It seems that the quest for a functioning molecular device based on metal-ion-modified DNA is about to begin. ■

Jens Müller is in the Department of Chemistry, University of Dortmund, Otto-Hahn-Straße 6, 44227 Dortmund, Germany.

e-mail: jens.mueller@uni-dortmund.de

1. Tanaka, K. *et al. Nature Nanotech.* **1**, 190–194 (2006).
2. Tanaka, K. *et al. Science* **299**, 1212–1213 (2003).
3. Clever, G. H., Polborn, K. & Carell, T. *Angew. Chem. Int. Edn* **44**, 7204–7208 (2005).
4. Meggers, E. *et al. J. Am. Chem. Soc.* **122**, 10714–10715 (2000).
5. Switzer, C. & Shin, D. *Chem. Commun.* 1342–1344 (2005).
6. Miyake, Y. *et al. J. Am. Chem. Soc.* **128**, 2172–2173 (2006).
7. Brotschi, C. & Leumann, C. J. *Nucleos. Nucleot. Nucl. Acids* **22**, 1195–1197 (2003).
8. Weizman, H. & Tor, Y. *J. Am. Chem. Soc.* **123**, 3375–3376 (2001).
9. Roelfes, G., Boersma, A. J. & Feringa, B. L. *Chem. Commun.* 635–637 (2006).
10. Porath, D., Cuniberti, G. & Di Felice, R. *Top. Curr. Chem.* **237**, 183–227 (2004).
11. Braun, E. *et al. Nature* **391**, 775–778 (1998).

ASTROPHYSICS

Unity among black holes

Jörn Wilms

Black holes box at two weights: active galactic nuclei are in the super-heavyweight class, whereas galactic black holes are relative featherweights. But does the same physics pack both objects' punches? It seems that it does.

The existence of black holes — agglomerations of mass so concentrated that their gravitational pull allows nothing, not even light, to escape from them — is perhaps the most intriguing prediction of Einstein's general theory of relativity. Since the early 1960s, astronomers have identified two types of compact source that are probably black holes. These are galactic black holes (GBH), which have masses between 5 and 20 times that of the Sun, and active galactic nuclei (AGN), which weigh in at many millions of solar masses. On page 730 of this issue, M^cHardy *et al.*¹ present observations of X-ray emissions from these sources that indicate that, despite the huge disparity in their masses, the physical mechanisms that power these two classes of object are the same.

Evidence for AGN is to be found at the centre of most galaxies, including our own Milky Way². AGN are point-like sources that radiate light equivalent to that of several billion stars — as much as their host galaxy. The maximum luminosity of GBH, on the other hand, is typically more modest, equivalent to that of a few tens of thousands of Sun-like stars. This radiation, it is important to note, does not emerge from the black hole itself. In both cases, the radiation is probably a by-product of a process known as accretion, in which matter from the host galaxy or a companion star is attracted towards the deep gravitational well of the black hole³.

A full understanding of accretion is one of the fundamental goals of astrophysics. The process is ubiquitous in the Universe, and many astrophysical objects have undergone episodes of accretion at some point in their life. The bare bones of the process are thought to be these: owing to conservation of angular momentum, material cannot fall straight into a compact, massive object such as a black hole, but forms a disk in which angular momentum is transported outwards as the accreted material spirals inwards⁴. At the disk's inner edge, the accreted gas reaches temperatures of 10^6 – 10^7 K, and so emits radiation at X-ray energies. The energy thus released is between 7% and 42% of the total energy of the gas at rest, making accretion onto a compact object the most efficient energy-producing process known. (For comparison, when hydrogen atoms fuse in the centre of our Sun, the energy released is just 0.7% of their rest energy.) Thus, despite their huge luminosity, AGN need to accrete material at a rate of just 1–2 solar masses a year to fuel themselves.

Because X-rays come from the hottest parts of the accretion disk, they provide diagnostic information about physics close to the black hole itself. AGN and GBH have similar X-ray spectra, so most astronomers believe that the physics of emission is similar in both. But secondary effects, such as absorption in the material surrounding the black hole, complicate comparison of AGN and GBH spectra, and direct proof for the two sources' similarity has been lacking.

M^cHardy *et al.*¹ study the time variability of AGN and GBH. The idea is that, if the physics of AGN and GBH emission is similar, the variability in the emission should be similar too. The typical timescales of the variation should, however, scale with mass⁵. GBH are aperiodically variable on timescales from days down to around 0.01 seconds, so similar variability patterns in AGN are expected on timescales of months to years. For this reason, their variability has been studied less, as measurements routinely performed in a few hours for GBH require observational campaigns of several years for AGN. Such measurements have become available only within the past decade through satellites such as NASA's Rossi X-ray Timing Explorer (RXTE).

To measure the contribution of variations at a given frequency, f , to a source's overall variability, astronomers use a mathematical transformation of the source's light curve known as its power spectral density. With black holes, the most common feature of this function is a break^{6,7} at which it turns from a shape roughly proportional to f^{-1} to one proportional to f^{-2} . Previous searches⁵ had shown that this break tends to occur at a lower frequency (on a slower timescale) the greater the mass of the object. But a large scatter in the data points remained, making it difficult to extrapolate from this apparent correlation any general conclusion on the similarity of the mechanisms underlying GBH and AGN.

M^cHardy *et al.*¹ show that the reason for this scatter is that the break frequency depends strongly on luminosity. Once the authors took this into account, they could describe the observed variability of ten AGN, extending over more than three magnitudes in mass accretion rate, with one relation. Extending the sample to GBH allowed the variability properties of these sources to be explained, too. It would be hard to credit that two different physical processes should show essentially the same timing behaviour when extrapolated over

such a large range of masses. Thus, these results provide a strong indication that the physics responsible for X-rays from stellar-mass and supermassive black holes is identical. The message is that for a full understanding of accretion onto compact, massive objects, objects of all sizes must be studied. At least here, the classical separation of astronomy into 'galactic', dealing with objects on the scale of GBH, and 'extragalactic', engaged with larger objects such as AGN, is artificial.

But with only ten AGN and two GBH, the number of objects in M^cHardy and colleagues' sample¹ is small. Further work is required to refine their relation by increasing the number of long-term AGN light-curves, and by adding data from more GBH-monitoring observations. Data from individual GBH should constrain the relationship between the characteristic timescale and the accretion rate: the shorter timescales of GBH mean that dramatic luminosity changes caused by changes in accretion can be observed in a single object⁸. Providing better data on AGN variability might be more challenging: understandably, it is difficult to convince the time-allocation committees of major satellites to commit themselves to such time-intensive long-term projects. RXTE is a commendable exception here.

For the next generation of satellites, dedicated, all-sky X-ray monitors are being discussed. These instruments will provide AGN light-curves without requiring repeated observations of single sources. Should these instruments come into being, together with dedicated instruments for the study of bright galactic sources, more intriguing insights should emerge from black holes. ■

Jörn Wilms is at the Dr. Karl Remeis-Sternwarte Bamberg, Astronomisches Institut der Universität Erlangen-Nürnberg, Sternwartstraße 7, 96049 Bamberg, Germany.

e-mail: joern.wilms@sternwarte.uni-erlangen.de

1. M^cHardy, I. M., Koeding, E., Knigge, C., Uttley, P. & Fender, R. P. *Nature* **444**, 730–732 (2006).
2. Schödel, R. *et al.* *Nature* **419**, 694–696 (2002).
3. Rees, M. *Annu. Rev. Astron. Astrophys.* **22**, 471–506 (1984).
4. Shakura, N. I. & Sunyaev, R. A. *Astron. Astrophys.* **24**, 337–355 (1973).
5. Edelson, R. & Nandra, K. *Astrophys. J.* **514**, 682–690 (1999).
6. Uttley, P., M^cHardy, I. M. & Papadakis, I. E. *Mon. Not. R. Astron. Soc.* **332**, 231–250 (2002).
7. Belloni, T. & Hasinger, G. *Astron. Astrophys.* **227**, L33–L36 (1990).
8. McClintock, J. E. & Remillard, R. A. in *Compact Stellar X-ray Sources* (eds Lewin, W. H. G. & van der Klis, M.) 157–214 (Cambridge Univ. Press, 2004).

Correction

In the News & Views article "Archaeology: High tech from Ancient Greece" by François Charette (*Nature* **444**, 551–552; 2006), there was a typographical slip in reference 5. It should read: Field, J. V. & Wright, M. T. *Ann. Sci.* **42**, 87–138 (1985). Also, we were a century out in dating the re-emergence of certain elements of the Antikythera Mechanism in Byzantium. According to current scholarship, this occurred in the sixth century, not the fifth as stated.

OBITUARY

Reinhart Heinrich (1946–2006)

Pioneer in systems biology.

In biology, mathematical systems analysis was until recently nearly invisible in the dazzling light of twentieth-century discoveries. But it has emerged from the shadows in the field of systems biology, a subject buoyed by immense data sets, conveyed by heavy computing power, and addressing seemingly incomprehensible forms of complexity. If systems biology has heroes, one of them is Reinhart Heinrich, a former professor at the Humboldt University in Berlin, who died on 23 October, aged 60. His most famous accomplishment was metabolic control theory, published in 1974 with Tom Rapoport and formulated independently by Henrik Kacser and James A. Burns in Edinburgh, UK.

From the 1930s to the 1960s, biochemists were busy describing metabolic pathways, just as molecular biologists today are feverishly trying to inventory the cell's gene-transcription and signalling circuits. The basic kinetic features of the enzymes in the major pathways were studied in great detail and with exceeding care. It seemed self-evident that, knowing the properties of each element, the behaviour of a pathway *in vivo* could simply be understood as the sum of its enzymes.

One assertion, drilled into the head of every biochemist, was the concept of the rate-limiting step. In this view, the flux through a pathway was determined by the slowest reaction, in the way a bucket brigade fighting a fire would be limited by the speed of the slowest member. Yet this concept, as shown by the metabolic control theory, was theoretically flawed, practically incomplete and often wrong, as many efforts to genetically engineer enzymes by changing 'rate-limiting steps' would ultimately show.

Metabolic control theory introduced the concept of control coefficients — dimensionless quantities indicating how the flux of a pathway depended on a given step. Only a pathway where every control coefficient except one was zero would have a rate-limiting step, since the flux of that pathway would depend only on that step. Several pathways in fact had rate-limiting steps, but that often reflected the structure of the pathway, and could not simply be deduced from the maximum rates of the individual enzymes, their Michaelis constants or their displacement from equilibrium. Many pathways were instead networks, with the fluxes distributed in a self-governing way among its various branches.

Heinrich went on to apply this theory to the real case of glycolysis in red blood cells,

where he showed that the flux of the reaction was shared by several enzymes. Much later, he extended his ideas to signal-transduction pathways, introducing control coefficients to dynamic processes. Sticking to real examples, such as the Wnt signalling and MAP kinase pathways, he again demonstrated that new properties and constraints emerge when the individual steps are combined into a complete pathway.

Heinrich also pointed the way to considerations of optimality theory and evolution that will confront systems biology for the next century. The question of evolution lies just beneath any effort to understand biology. Yet in most cases, physiological function and evolutionary change are considered distinct and are investigated by different people. Heinrich's work illustrated how systems biology might develop, where the central question will be not only how a system works, but why it has the specific design it has. In examining optimization for metabolic pathways, he found that there is no unique solution for the individual rate constants and affinities, but many solutions that are each optimized with respect to certain conditions. By considering the criteria for optimization, such as pathway flux or minimization of enzyme concentration, he nevertheless found general tendencies, such as a match between the Michaelis constant and substrate concentration. Evolutionary optimization is becoming a central feature in systems biology.

Some of the wrenching political developments of the twentieth century framed Heinrich's life, and it is a tribute to the collegiality of science that his reputation and influence were able to transcend such considerations. He was born in Dresden, but spent his early years in Kuibishev, Russia. His father, a professor of applied mathematics, belonged to the group of scientists in East Germany who, after the Second World War, were rounded up by the Russians to help build their aviation industry.

The family returned to East Germany in the early 1950s, and Heinrich studied solid-state physics in Dresden. But after his PhD he became interested in applying physics to biology, and moved to East Berlin to join the group of Samuel Rapoport — a family connection that later continued when he



worked with the son, Tom. Rapoport senior was a biochemist who himself was a refugee (in this case from anti-communist hysteria in the United States). In East Berlin, travel and outside communication were severely restricted, but Heinrich partly coped with the isolation by organizing meetings in East Germany. When the Berlin Wall fell, his fame began to spread more rapidly, increasing as systems biology has taken off as an organized discipline over the past few years.

Reinhart Heinrich's character is revealed both in his originality in seeking rational explanations in biology and in his approach to life in general. He published a prizewinning philosophical novel (*Jenseits von Babel* [*Beyond Babel*]) and a book of poems. He played the violin and was well versed in history and literature; he also had an amazing ability to remember what he had done on every day of his life. His humour and charm were contagious. He trained a whole generation of students, becoming the father of the Berlin school of theoretical biophysics, and led us to the edge of a new domain of knowledge. Reinhart died too young. But he was fortunate enough to have glimpsed the future outlines and promise of systems biology, a field he himself was instrumental in founding.

Marc W. Kirschner

Marc W. Kirschner is at Harvard Medical School, Boston, Massachusetts 02115, USA.
e-mail: marc@hms.harvard.edu

BRIEF COMMUNICATIONS

Nectar bat stows huge tongue in its rib cage

The extreme length of this bat's tongue might have coevolved with the long flowers it pollinates.

Bats of the subfamily Glossophaginae (family Phyllostomidae) are arguably the most specialized of mammalian nectarivores, and hundreds of neotropical plants rely on them for pollination^{1,2}. But flowers pollinated by bats are not known to specialize for bat subgroups (unlike flowers that have adapted to the length and curvature of hummingbird bills, for example), possibly because the mouthparts of bats do not vary much compared with the bills of birds or the probosces of insects^{3,4}. Here I report a spectacular exception: a recently-described nectar bat that can extend its tongue twice as far as those of related bats and is the sole pollinator of a plant with corolla tubes of matching length.

The nectar bat *Anoura fistulata* was discovered in the cloud forests of the Andes of Ecuador, where it lives with two other glossophagines, *A. caudifer* and *A. geoffroyi*⁵. I measured the tongues of these bats by placing them in separate flight cages and training them to drink sugared water from a modified straw (6 mm in diameter; see supplementary information). The *A. caudifer* drank to a maximum depth of 36.7 mm (± 2.28 s.d., $n = 6$) and the *A. geoffroyi* to 39.3 mm (± 2.21 s.d., $n = 5$). The tongue of the *A. fistulata*, however, reached 84.9 mm (± 3.51 s.d., $n = 4$), which is 150% of its body length (Fig. 1a). This relative tongue protrusion is greater than that seen in any other mammal, and is second only to the chameleon⁶ among the vertebrates.

Across the glossophagine nectar bats, maximum tongue extension is tightly correlated with the length of their rostral components, such as the palate and mandible⁷. Although the correlation holds for *A. caudifer* and *A. geoffroyi*, *A. fistulata* falls far outside the 95% confidence interval (Fig. 1b). Close examination of tongue morphology reveals the basis for this pattern. In other nectar bats, the base of the tongue coincides with the base of the oral cavity⁸ (the typical condition for mammals), but in *A. fistulata* the tongue passes back through the neck and into the thoracic cavity. This portion is surrounded by a sleeve of tissue, or glossal tube, which follows the ventral surface of the trachea back and positions the base of the tongue between the heart and the sternum (Fig. 1c).

Darwin famously (and correctly) predicted that the long-spurred Malagasy star orchid (*Angraecum sesquipedale*) would be pollinated by a hawkmoth with an extremely long proboscis⁹. Likewise, a nectar bat that has such extreme tongue extension points to the existence of a flower that has a corolla of comparable length.

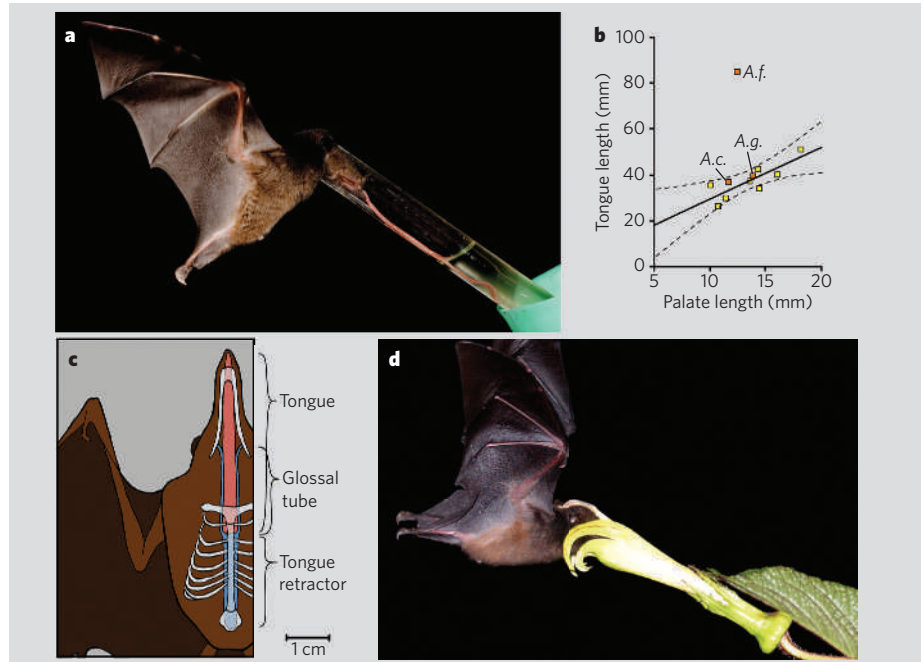


Figure 1 | Protrusible tongue of the nectar bat *Anoura fistulata*. a, *Anoura fistulata* feeding from a test tube filled with sugared water; its tongue (pink) can extend to 150% of body length. b, Relation between maximum tongue extension and palate length for 11 species of glossophagine nectar bat. *Anoura fistulata* (*A. f.*) is an outlier; solid line, regression for ten other bats ($y = 2.25x + 6.78$, $r^2 = 0.68$); dashed lines, 95% confidence intervals. Orange squares, *A. fistulata*, *A. geoffroyi* and *A. caudifer*; yellow squares, data for species from ref. 7 (from left, *Lichonycteris obscura*, *Glossophaga commissarisi*, *G. soricina*, *Hylonycteris underwoodi*, *A. cultrata*, *Lonchophylla robusta*, *Leptonycteris curosoae*, *Choeronycteris mexicana*). c, Ventral view of *A. fistulata*, showing tongue (pink), glossal tube and tongue retractor muscle (blue), and skeletal elements (white). d, *Anoura fistulata* pollinating the specialized flower of *Centropogon nigricans*; because of the long corolla, only *A. fistulata* can reach its nectar. (Fig. 1a, M. Cooper; Fig. 1d, N. M.)

Dietary studies of *Anoura* in four reserves on the eastern and western slopes of the Andes confirm this prediction. During 129 nights of mist-netting in 2003–05, I captured 46 *A. geoffroyi*, 38 *A. caudifer*, and 21 *A. fistulata*, and identified the pollen on their fur and in their faeces. Pollen from *Centropogon nigricans*, which has corollas 8–9 cm long, was carried only by *A. fistulata* (Fig. 1d), as might be expected, given that other *Anoura* could not reach this flower's nectar. During 55 hours of nocturnal and diurnal videotaping of 12 flowers of *C. nigricans*, ten bats were the only visitors. This observation, combined with the finding that *A. fistulata* was the only bat visitor, supports the conclusion that *A. fistulata* is the only pollinator of this plant.

Specialization on one species of pollinator is exceedingly rare in angiosperms¹⁰, and *C. nigricans* is the only example known in flowers pollinated by bats. After the initial evolution of

a glossal tube, the extreme tongue length of *A. fistulata* probably coevolved with long flowers such as those of *C. nigricans*. In an example of convergent evolution, pangolins (scaly anteaters) also have a glossal tube¹¹; despite their different diets, ant-eating and nectar-feeding animals face similar evolutionary pressures for highly protrusible tongues, and pangolins and *A. fistulata* have independently converged on a similar solution.

Nathan Muchhala

Department of Biology, University of Miami, Coral Gables, Florida 33143, USA
e-mail: muchhala@bio.miami.edu

- von Helversen, O. & Winter, Y. in *Bat Ecology* (eds Kunz, T. H. & Fenton, M. B.) 346–397 (University of Chicago Press, Chicago, 2003).
- Dobart, K. & Peikert-Holle, T. *Blüten und Fledermäuse* (Waldemar Kramer, Frankfurt am Main, 1985).
- Stiles, F. G. *Am. Zool.* **18**, 715–727 (1978).
- Fleming, T. H., Muchhala, N. & Ornelas, J. F. in *Contribuciones Mastozoológicas en Homenaje a Bernardo*

- Villa (eds Sánchez-Cordero, V. & Medellín, R. A.) 163–186 (Instituto de Biología e Instituto de Ecología, UNAM, Mexico, 2005).
5. Muchhala, N., Mena, P. V. & Albuja, L. V. *J. Mammal.* **86**, 457–461 (2005).
6. Herrel, A., Meyers, J. J., Aerts, P. & Nishikawa, K. C. *J. Exp. Biol.* **204**, 3621–3627 (2001).

7. Winter, Y. & von Helversen, O. *J. Mammal.* **84**, 886–896 (2003).
8. Griffiths, T. A. *Am. Mus. Nov.* **2742**, 1–45 (1982).
9. Darwin, C. *On the Various Contrivances by which British and Foreign Orchids are Fertilized* (Murray, London, 1862).
10. Fenster, C. B. et al. *Annu. Rev. Ecol. Evol. Sys.* **35**, 375–403 (2004).

11. Chan, L. K. *J. Mammal.* **76**, 472–480 (1995).

Supplementary information accompanies this communication on Nature's website.

Received 13 September; accepted 10 November 2006.

Competing financial interests: declared none.
doi:10.1038/444701a

NAVIGATION

Bat orientation using Earth's magnetic field

Bats famously orientate at night by echolocation¹, but this works over only a short range, and little is known about how they navigate over longer distances². Here we show that the homing behaviour of *Eptesicus fuscus*, known as the big brown bat, can be altered by artificially shifting the Earth's magnetic field, indicating that these bats rely on a magnetic compass to return to their home roost. This finding adds to the impressive array of sensory abilities possessed by this animal for navigation in the dark.

For some taxa, navigation behaviour can be readily investigated in the laboratory³. To study the wide-ranging navigation of bats, however, their flight path needs to be tracked in a natural setting. Limitations of the available technology make this a labour-intensive process, so bat navigation is relatively poorly understood compared with that of other animals².

We used radio telemetry from a small aircraft⁴ to track big brown bats displaced 20 km north of their home roost (for methods, see supplementary information). A control group released from this site headed in a direction significantly towards home (see supplementary information) at 5 km from the release site (Fig. 1a).

To test whether bats use the Earth's magnetic field, we exposed two groups of bats to a rotated magnetic field, one 90° clockwise and one 90° anticlockwise with respect to magnetic north, for a period from 45 min before to 45 min after

sunset. This also allowed us to test whether the Earth's magnetic field was being used in conjunction with other cues, such as the sunset or stars⁵ (see supplementary information).

The headings of the clockwise group were significantly oriented in an easterly direction (90°) at 5 km from the release site, whereas the anticlockwise group moved significantly in a westerly (270°) direction; the two groups showed a significant difference (Fig. 1a). These different initial orientations of the groups indicate that they may have been using a sunset-calibrated magnetic compass^{5,6}.

Some experimental bats corrected and homed during the same night, despite being initially orientated away from home (Fig. 1b, c). Although such behaviour has previously been unknown in bats, homing pigeons can correct and return home after an initial deviation when clock-shifted⁷. We suggest that the deflected bats that nevertheless home during the same night recognize a mismatch between the direction they are flying and their navigational map.

Besides the application described here to measure bat navigation, radio telemetry has also been used to investigate migration in insects⁸. The possibility of transmitting such radio signals to low orbiting satellites should open up field studies on the orientation, navigation and migration of small, wide-ranging animals.

Richard A. Holland*†, **Kasper Thorup***‡§, **Maarten J. Vohhof**||, **William W. Cochran**¶, **Martin Wikelski***

*Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey 08544, USA

e-mail: rahollan@princeton.edu

†Institute of Integrative and Comparative Biology, University of Leeds, Leeds LS2 9JT, UK

‡Zoological Museum, and §Center for Macroecology, Institute of Biology, University of Copenhagen, 2100 Copenhagen, Denmark

||Department of Biological Sciences, Western Michigan University, Kalamazoo, Michigan 49008, USA

¶Illinois Natural History Survey, Champaign-Urbana, Illinois 61820, USA

1. Simmons, J. A. et al. in *Hearing by Bats* (eds Popper, A. N. & Fay, R. R.) 146–190 (Springer, London, 1995).
2. Neuweiler, G. *The Biology of Bats* (Oxford University Press, New York, 2000).
3. Berthold, P. *Bird Migration: A General Survey* (Oxford University Press, Oxford, 2001).
4. Michener, M. C. & Walcott, C. *J. Exp. Biol.* **47**, 99–131 (1967).
5. Cochran, W. W., Mouritsen, H. & Wikelski, M. *Science* **304**, 405–408 (2004).
6. Muheim, R., Moore, F. R. & Phillips, J. B. *J. Exp. Biol.* **209**, 2–17 (2006).
7. Bonadonna, F., Holland, R., Dall'Antonia, L., Guilford, T. & Benvenuti, S. *J. Exp. Biol.* **203**, 207–212 (2000).
8. Wikelski, M. et al. *Biol. Lett.* **2**, 325–329 (2006).

Supplementary information accompanies this communication on Nature's website.

Received 24 July; accepted 25 October 2006.

Competing financial interests: declared none.
doi:10.1038/444702b

BRIEF COMMUNICATIONS ARISING online
♦ www.nature.com/bca see Nature contents.

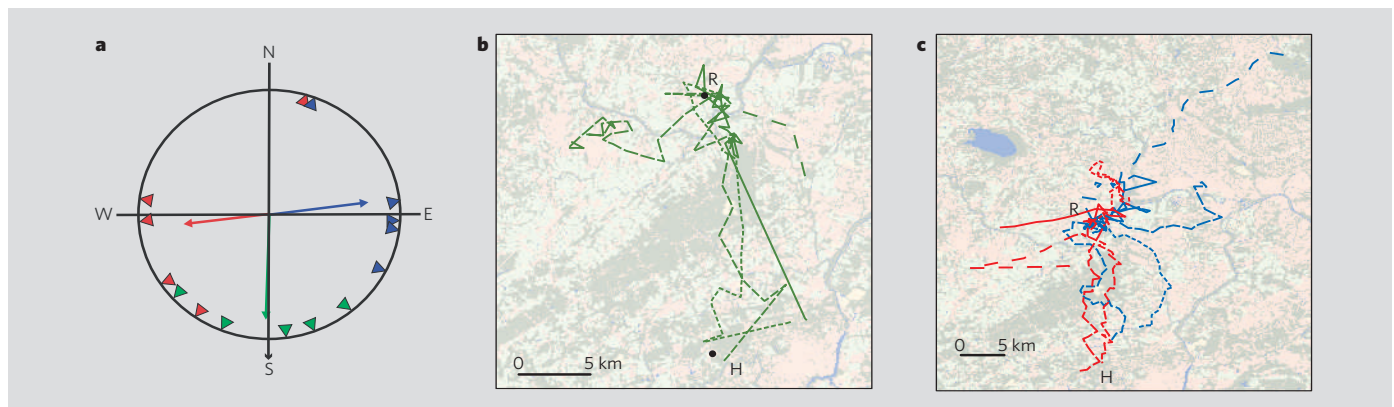


Figure 1 | Headings and tracks of homing bats. **a**, Heading directions at 5 km after release 20 km north of the home roost (to south; black arrow). Arrowheads, directions for individual bats; arrows, mean direction for the group. Red, anticlockwise (ACW) rotation of magnetic field by 90° with respect to north; blue, clockwise (CW) rotation of magnetic field by 90°; green, controls (no rotation of magnetic field). Orientation was significantly southerly in controls (V test, 180°, $U = 2.862$, $P = 0.0072$); westerly in ACW bats (V test, 270°, $U = 1.973$, $P = 0.023$); and easterly in CW bats (V test, 90°, $U = 2.66$, $P = 0.002$). Headings differed significantly between the three groups (Watson–Williams 3-sample test: $F = 16.808$, $P = 0.00033$; pairwise: CW vs control, $F = 23.774$, $P = 0.001$; ACW vs control, $F = 6.733$, $P = 0.032$; ACW vs CW, $F = 23.503$, $P = 0.001$). **b**, **c**, Control (**b**) and experimental tracks (**c**) of bats, with different dotted and dashed lines for individual bats ($n = 5$ in each group). Colours indicate direction of rotation of magnetic field, as in **a**. R, release site; H, home.

EVOLUTIONARY BIOLOGY

Evidence for sympatric speciation?

Arising from: M. Barluenga, K. N. Stölting, W. Salzburger, M. Muschick & A. Meyer *Nature* 439, 719–723 (2006)

Sympatric speciation is difficult to demonstrate in nature and remains a hotly debated issue. Barluenga *et al.*¹ present a case of putative sympatric speciation for two cichlid species in the Nicaraguan crater lake Apoyo, but they overlook or reinterpret some key published information on the system. Although sympatric speciation is possible in theory^{2,3}, we show here that, when this information is taken into account, the results of Barluenga *et al.*¹ do not provide conclusive evidence for sympatric speciation: this is because the null hypothesis of multiple invasion with introgression cannot be rejected.

Figure 2a in Barluenga *et al.*¹ is a modified version of Fig. 3 in Barluenga and Meyer⁴ (compared in Fig. 1), in which the central mitochondrial (mt) DNA haplotype, from which the Lake Apoyo haplotypes are derived, is shared by two species from Lake Nicaragua, *Amphilophus labiatus* and *A. citrinellus*, at similar frequencies, and is also found in different members of the *A. citrinellus* complex from other lakes. These results therefore do not allow any conclusion about which of these species is ancestral to the mitochondrial lineage of Lake Apoyo cichlids. It follows that the restriction by Barluenga *et al.*¹ of their subsequent analyses of nuclear DNA markers to only *A. citrinellus* is unjustified.

Mitochondrial monophyly of the Lake Apoyo specimens, as indicated by the mtDNA haplotype network of Barluenga *et al.*¹ (which is based on a single point mutation), does not imply monophyly of a species pair that can exchange genes. Multiple colonization with introgression is frequently associated with mitochondrial monophyly too, as shown for sympatric stickleback species and other sympatric fish^{5–7}, and in fact the nuclear genetic data of Barluenga *et al.*¹ support this scenario. In all analyses of variation at selectively neutral nuclear genes, *A. citrinellus* of Lake Apoyo is much closer to *A. citrinellus* of Lake Nicaragua than is *A. zaliosus*. This seems incompatible with sympatric speciation from one founder, which, by definition, leaves both daughter species equally related to their common ancestor when measured over many neutral loci.

Their factorial correspondence analysis (FCA) shows that *A. citrinellus* from Lake Apoyo is genetically intermediate between *A. citrinellus* from Lake Nicaragua and *A. zaliosus* from Lake Apoyo (see Fig. 2d of ref. 1); this is exactly as would be predicted if the Lake Apoyo population of *A. citrinellus* was the result of a second wave of colonization from Lake Nicaragua that was followed by introgression from the popula-

tion that had resulted from the first colonization — namely *A. zaliosus*. Consistent with this, their FCA and bayesian population-assignment tests both reveal nuclear genotypes from Lake Nicaragua and admixed genotypes in the sample of *A. citrinellus* from Lake Apoyo but not in *A. zaliosus*, and FCA and bayesian population-assignment tests indicate introgression between these two species (Fig. 2d and Fig. 3 in ref. 1).

The claim by Barluenga *et al.*¹ of monophyly of Lake Apoyo cichlids is also weakened by their incomplete sampling of all the species in the lake. This lake contains four or five distinct and assortatively mating *Amphilophus* forms, referred to as *A. cf. citrinellus*, *A. zaliosus*, *A. “Apoyo amarillo”*, *A. “chancho”* and *A. “short”*, for which genetic and phenotypic information is available⁸, although they have not been formally described. Barluenga *et al.*¹ acknowledge this work⁸ but mention only two species, rather than the four or five that are found in Lake Apoyo; also, several other described species of the *A. citrinellus* species complex are known to exist in the area⁹. Recognition and analysis of the species-level diversity of the system is a prerequisite for reconstructing the evolutionary history of speciation. However, it is unclear whether any of these other species

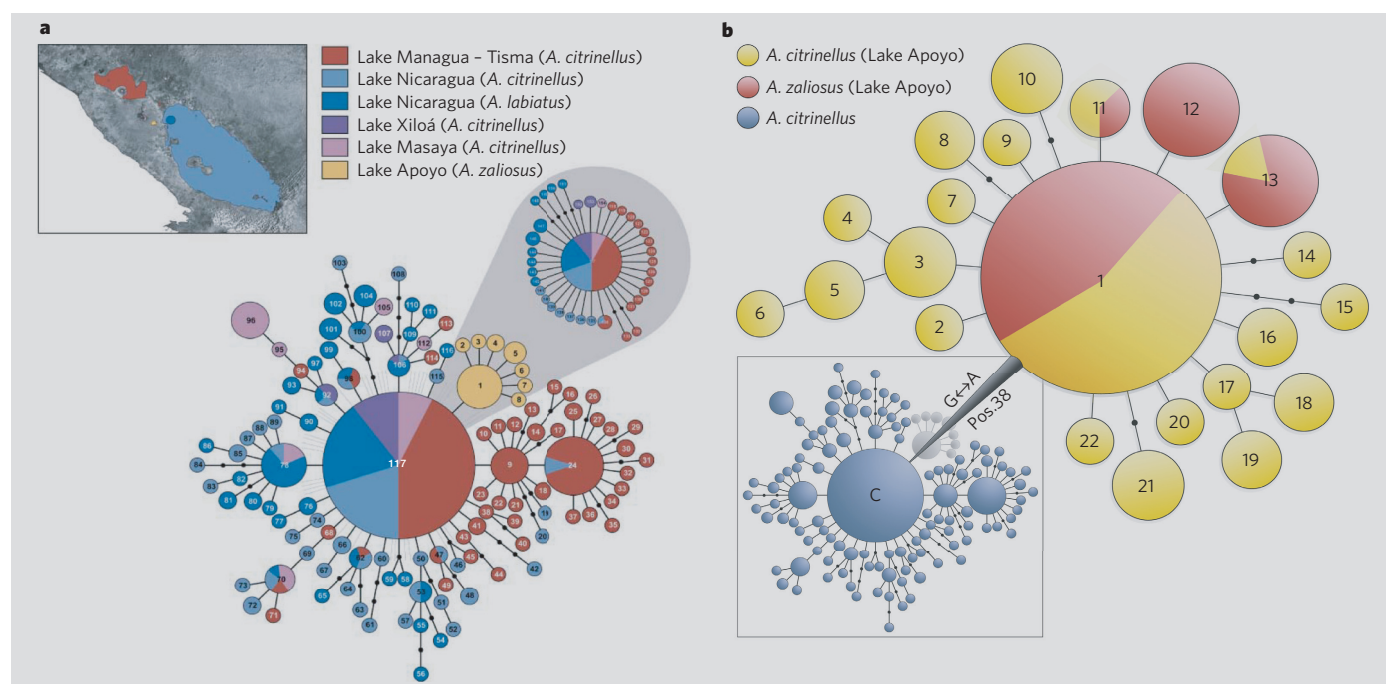


Figure 1 | Contrasting information in two identical haplotype networks of Nicaragua cichlids. a, The mitochondrial haplotype ancestral to Lake Apoyo (haplotype 117 in Fig. 3 of Barluenga *et al.*⁴; reproduced with permission) is shared between *Amphilophus citrinellus* from five different lakes and *A. labiatus* from Lake Nicaragua. **b**, Haplotype C in Fig. 2a of Barluenga *et al.*¹ is identical to the haplotype 117 in **a**. The central haplotype networks (blue in ref. 1; blue, pink and red in ref. 4) are identical, except that all haplotypes assigned to *A. labiatus* in ref. 4 have been assigned to *A. citrinellus* in ref. 1. Note also the difference in *A. zaliosus* haplotype counts (four versus eight) in the two figures.

are included in the *A. cf. citrinellus* sample of Barluenga *et al.*¹. Their morphological analysis does not justify the authors' conclusions about the number of morphologically differentiated taxa, because even *A. zaliosus* and *A. citrinellus* broadly overlap in morphospace (Fig. 4b of ref. 1).

Because Barluenga *et al.*¹ exclude *A. labiatus* and overlook the phenotypic and taxonomical complexity of the *A. citrinellus* complex, their microsatellite-based phylogenetic inferences (see their supplementary Fig. 3a, b) cannot show monophyly of their two species in Lake Apoyo. These phylograms are based on allele frequencies for which the authors have simply pooled samples into *A. zaliosus* and *A. cf. citrinellus* "Apoyo".

In conclusion, the intermediate nuclear-genetic position of the Lake Apoyo *A. citrinellus* population between *A. zaliosus* and *A. citrinellus* from Lake Nicaragua is incompatible with sympatric speciation. Instead, it indicates that two invasions occurred, followed by introgressive hybridization and fixation of one mitochondrial

haplotype — as in other fish species⁷. The close proximity of Lake Apoyo and Lake Nicaragua makes this easily possible.

Because *A. citrinellus* and *A. labiatus* in Lake Nicaragua are hardly distinguishable at microsatellite loci⁴ and their mtDNA sequences are indistinguishable, we do not yet know whether these colonizations involved two waves of one of these species, or one of each. It will be necessary to test these alternatives, and to determine whether genetic similarity of the Lake Apoyo endemics is due to secondary introgression or shared ancestry^{7,10,11}.

U. K. Schlieven*, T. D. Kocher†, K. R. McKaye‡, O. Seehausen§||, D. Tautz¶

* Bavarian State Collection of Zoology, 81247 München, Germany

e-mail: schlieven@zsm.mwn.de

† Hubbard Center for Genome Studies, University of New Hampshire, Durham, New Hampshire 03824, USA

‡ Appalachian Laboratory, UMCES, Frostburg, Maryland 21532, USA

§ Aquatic Ecology & Evolution, Institute of Zoology,

University of Bern, 3012 Bern, Switzerland

|| EAWAG Ecology Center, 6047 Kastanienbaum, Switzerland

¶ Department of Evolutionary Genetics, Institute of Genetics, University of Cologne, 50674 Cologne, Germany

1. Barluenga, M., Stölting, K. N., Salzburger, W., Muschick, M. & Meyer, A. *Nature* **439**, 719–723 (2006).
2. Dieckmann, U., Doebeli, M., Metz, H. & Tautz, D. *Adaptive Speciation* (Cambridge Univ. Press, Cambridge, UK, 2004).
3. Gavrilov, S. *Fitness Landscapes and the Origin of Species*. Monographs in Population Biology 41 (Princeton University Press, Princeton, New Jersey, 2004).
4. Barluenga, M. & Meyer, A. *Mol. Ecol.* **13**, 2061–2076 (2004).
5. Taylor, E. B. & McPhail, J. D. *Proc. R. Soc. Lond. B* **267**, 2375–2384 (2000).
6. Bernatchez, L. *et al.* *Can. J. Fish. Aqu. Sci.* **52**, 179–185 (1995).
7. Seehausen, O. *Trends Evol. Ecol.* **19**, 198–207 (2004).
8. McKaye, K. R. *et al.* *Cuadernos de Investigación de la UCA* **12**, 19–47 (2002).
9. Stauffer, J. R. & McKaye, K. R. & Stauffer, J. R. *Cuadernos de Investigación de la UCA* **12**, 1–18 (2002).
10. Machado, C. A. *et al.* *Mol. Biol. Evol.* **19**, 472–488 (2002).
11. Schlieven, U. K. & Klee, B. *Front. Zool.* **1**, doi:10.1186/1742-9994-1-5 (2004).

doi:10.1038/nature05419

EVOLUTIONARY BIOLOGY

Barluenga *et al.* reply

Replying to: U. K. Schlieven, T. D. Kocher, K. R. McKaye, O. Seehausen & D. Tautz *Nature* **444**, doi:10.1038/nature05419 (2006)

We reported a case of sympatric speciation in the Nicaraguan Midas cichlid species complex¹. Schlieven *et al.*² question the interpretation of aspects of our data, but their proposed alternative scenario of multiple colonization and hybridization is considerably less parsimonious, contains some inconsistencies, and is incompatible with the available evidence.

Amphilophus labiatus is not a sister species of the Lake Apoyo *Amphilophus* fauna³. The central haplotype in Fig. 2 of ref. 1 indeed contains specimens of *A. labiatus* and *A. citrinellus*; this figure, as indicated¹, is a simplified version of our earlier one³. However, we have shown that *A. labiatus* is more distantly related to the monophyletic Lake Apoyo assemblage than is *A. citrinellus* from Lake Nicaragua³ (*Nature* did not permit us to show additional analyses or figures to this effect). This is also supported by morphometrics⁴ and the absence of *A. labiatus* from Lake Apoyo. Our microsatellite, mitochondrial (mt) DNA and amplified fragment-length polymorphism (AFLP) analyses¹ confirm that *A. zaliosus* and *A. citrinellus* from Lake Apoyo are each other's closest relatives.

There is no evidence to support the assertion by Schlieven *et al.*² that *A. citrinellus* of Lake Apoyo is closer to *A. citrinellus* of Lake Nicaragua than is *A. zaliosus*. Factorial correspondence analysis does not either, as it illustrates present but not past genetic distances (in fact, any ancestral population should be equidistant from all of its descendants). Similarly,

with only three potential cases in the more than 120 individuals included from Lake Apoyo (as determined by the analyses using Structure software; see Fig 3 in ref. 1), introgression is very rare in *A. citrinellus* — if it exists at all ($P < 80\%$), as determined by the Structure analysis. The argument by Schlieven *et al.*² for secondary introgression from Lake Nicaragua into Lake Apoyo is based on a single specimen, which is unlikely to be an introgressant as it contains alleles of the genomes of all three populations, which is likely to be an artefact of the analysis. The monophyly of Lake Apoyo's *Amphilophus* species and the complete endemism of its mtDNA haplotypes argue against secondary colonization.

Instead, the analyses all indicate that *A. zaliosus* evolved sympatrically from *A. citrinellus* within Lake Apoyo. We showed that *A. zaliosus* is only about half as old as *A. citrinellus* from Lake Apoyo¹ (note that *A. citrinellus* from Lake Apoyo carries only a subset of the global *A. citrinellus* microsatellite alleles and that *A. zaliosus* carries only about half of the Lake Apoyo *A. citrinellus* alleles). These genetic data therefore rule out the alternative scenario proposed by Schlieven *et al.*, in which *A. citrinellus* entered Lake Apoyo in a second wave of colonization after *A. zaliosus*.

We do not believe that our sampling of the taxonomic diversity in Lake Apoyo was inadequate. Our Lake Apoyo data set does include morphs that others⁵ call "chanchó", "short" and

"amarillo". These morphotypes have never been formally described as species, no voucher specimens and no phenotypic or meristic information is available (only some photographs), and no experimental or observational data have been published that would support assortative mating. A previous genetic analysis⁵ based on three microsatellites yielded inconclusive results. Our own, much more detailed, analyses¹ find, so far, evidence for only two genetically discernable units of *Amphilophus* in Lake Apoyo — *A. zaliosus* and *A. citrinellus*. The seeming overlap in morphospace between the two Lake Apoyo species is due only to the two-dimensional projection of a multidimensional plot.

In summary, we maintain that the data fully support our original interpretations, whereas Schlieven *et al.*² propose a much less likely scenario that is not supported by the available data.

Marta Barluenga*†, Kai N. Stölting*†, Walter Salzburger*†, Moritz Muschick*, Axel Meyer*

*Lehrstuhl für Zoologie und Evolutionsbiologie, Department of Biology, University of Konstanz, 78457 Konstanz, Germany

e-mail: axel.meyer@uni-konstanz.de

†Present addresses: Zoologisches Museum, University Zürich, 8057 Zürich, Switzerland (K.N.S.); Department of Ecology and Evolution, University of Lausanne, Le Biophore, 1015 Lausanne, Switzerland (M.B., W.S.)

1. Barluenga, M., Stölting, K. N., Salzburger, W., Muschick, M. & Meyer, A. *Nature* **439**, 719–723 (2006).
2. Schlieven, U. K., Kocher, T. D., McKaye, K. R., Seehausen, O. & Tautz, D. *Nature* **444**, doi:10.1038/nature05419 (2006).
3. Barluenga, M. & Meyer, A. *Mol. Ecol.* **13**, 2061–2076 (2004).
4. Klingenberg, C. P., Barluenga, M. & Meyer, A. *Biol. J. Linn. Soc.* **80**, 397–408 (2003).
5. McKaye, K. R. *et al.* *Cuadernos de Investigación de la UCA* **12**, 19–47 (2002).

doi:10.1038/nature05420

CONTINENTAL RUNOFF

A quality-controlled global runoff data set

Arising from: N. Gedney et al. *Nature* **439**, 835–838 (2006)

Gedney *et al.*¹ attribute an increase in the twentieth-century continental runoff to the suppression of plant transpiration by CO₂-induced stomatal closure, by replicating a continental runoff data set². However, we have concerns about this data set and the methods used to construct it, in addition to those already raised³, which we believe may undermine their conclusions.

The continental runoff data set covers the period 1875–1994 and the start and end dates of runoff observations are listed² for each of 221 constituent stations. A wavelet-based runoff ‘reconstruction’ methodology² is used to infill any missing data and extend records forwards (to 1994) and backwards (to 1875) where necessary. In addition to previous concerns³ about the use of runoff stations subject to confounding anthropogenic influences, such as reservoirs, and aspects of the runoff reconstruction methodology, it seems that the runoff reconstruction methodology is based on the best correlation against one of ten unspecified reference stations². There is no discussion about whether these reference stations are spatially and temporally representative of the full range of runoff regimes being reconstructed around the world and not subject to confounding anthropogenic

influences. The sensitivity of the final reconstructed runoff to the choice of these reference stations is not properly addressed^{2,4}.

A fundamental requirement of a runoff data set used in an attribution study, such as that by Gedney *et al.*¹, is that it is representative of the observed runoff conditions around the world. The concerns outlined above and previously³ about the reconstruction methodology call into question the degree to which the runoff data set used² is representative of those conditions. Key to this is a discussion of the percentage of continental runoff that is reconstructed (infilled or extrapolated), which is not presented^{1,2,4}. Based on the start and end dates of the runoff records², extrapolation from 10–20 years of observed runoff data to the centennial scale⁴ has been applied to at least 34% of the stations. This is likely to be an underestimate, because not all stations have complete records between these dates.

Gedney *et al.*¹ analyse continental runoff records based on “observations from at least 20% of the total river basin area”, which could be taken to mean that up to 80% of the continental runoff is reconstructed, not observed, for some periods of their analysis. Without knowing the degree to which the runoff data

set is reconstructed and therefore representative, conclusions based on replication of that runoff data in a modelling (attribution, for example) analysis are speculative.

Considering the serious concerns about the continental runoff data set underlying the results of Gedney *et al.*¹, their conclusion that an increase in twentieth-century continental runoff is attributable to the suppression of plant transpiration by CO₂-induced stomatal closure is called into question. This highlights the need for the development of a quality-controlled, freely available, global runoff data set, which can be used with confidence (or at least informed caution) in future studies.

Murray C. Peel, Thomas A. McMahon

Department of Civil & Environmental Engineering, University of Melbourne, Victoria, 3010, Australia
e-mail: mpeel@unimelb.edu.au

1. Gedney, N. *et al.* *Nature* **439**, 835–838 (2006).
2. Labat, D., Godd  ris, Y., Probst, J. L. & Guyot, J. L. *Adv. Water Resour.* **27**, 631–642 (2004).
3. Legates, D. R., Lins, H. F. & McCabe, G. J. *Adv. Water Resour.* **28**, 1310–1315 (2005).
4. Labat, D., Godd  ris, Y., Probst, J. L. & Guyot, J. L. *Adv. Water Resour.* **28**, 1316–1319 (2005).

doi:10.1038/nature05480

CONTINENTAL RUNOFF

Gedney et al. reply

Replying to: M. C. Peel & T. A. McMahon *Nature* **444**, doi: 10.1038/nature05480 (2006)

Peel and McMahon¹ argue that limitations of the continental runoff data set² we use in our study³ call our main conclusion into question — that is, that changes in twentieth-century continental-scale runoff are partly attributable to the suppression of CO₂-induced stomatal closure. We support Peel and McMahon’s pleas for better and more complete runoff data, but we remain confident in our conclusions, as explained here.

First, we do not rely on the conclusion of Labat *et al.*² that there is a “link between global warming and the intensification of the global hydrological cycle” being correct, which has been questioned⁴. As we are accounting for anthropogenic influences in our study³, their presence in basins used in the runoff data set is also not an issue here.

Second, our statistical analysis includes data uncertainty. We use a restricted maximum-likelihood technique that takes account of temporal correlation in the uncertainty, as we

expect the data-set errors to be time-dependent. This produces wider and more realistic error bars than standard fitting techniques. As we are modelling timescales from annual to multi-decadal, our ability to reproduce the runoff data set at such a wide spread of temporal scales acts as a strong crosscheck. Yet, we still find that the effects of CO₂-induced stomatal closure are detectable at a statistically significant level.

Third, we find that, over every continent except Europe, including the direct CO₂ effect produces temporal behaviour that is closer to the observational data than when only the climate effect is included. This is also the case over Europe if we incorporate the net effect of human water consumption^{5,6}. Of the non-climatic mechanisms considered, only the direct CO₂ effect has both the temporal and spatial structure that is consistent with the observational runoff². The likelihood that such a signature is in the observational data set by

chance is therefore small, given that any errors in the data set² are likely to vary spatially and temporally.

For the reasons given³, we consider periods when the observed runoff areal coverage is at least 20% of the total, thereby avoiding some of the interpolation and gap-filling problems referred to by Peel and McMahon¹. In addition, as already mentioned³ and as discussed in more detail here, our results are consistent when we consider different time periods. The proportion of areal observation coverage in the runoff data set at a given year can be calculated from tables in ref. 2. However, the total observed discharge is likely to be more relevant for assessing the influence of reconstructed data². Indeed, we can obtain a good estimate of the proportion of observed runoff within the data set² by considering readily available measurements (Global Runoff Data Centre, Koblenz, Germany) from those rivers with large annual discharges (more than 6,000 m³ s^{−1}). The total runoff of this subset accounts for about 80% of gauged runoff total in the data set², making it a reasonable representation. In this sub-sample, 63–92% of the observations are available over the 1950–95 period.

As already mentioned³, our results are actu-

ally insensitive to the exact time periods used. For example, if we consider start times ranging from 1901 (when there is a low percentage of observed areal coverage and total discharge²) to 1950 (by which time most of the data² are from observations), we obtain the same results. Our conclusions are therefore insensitive to the exact coverage of the observed runoff data.

N. Gedney^{*†}, **P. M. Cox**[‡], **R. A. Betts**[§],
O. Boucher[§], **C. Huntingford**[†], **P. A. Stott**^{||}

^{*}Met Office, Hadley Centre for Climate Prediction

and Research (JCHMR), and [†]Centre for Ecology and Hydrology, Maclean Building, Wallingford OX10 8BB, UK

e-mail: nicola.gedney@metoffice.gov.uk

[‡]Mathematics Research Institute, University of Exeter, Exeter EX4 4QF, UK

[§]Met Office, Hadley Centre for Climate Prediction and Research, Exeter EX1 3PB, UK

^{||}Met Office, Hadley Centre for Climate Prediction and Research (Reading Unit), Meteorology Building, University of Reading, Earley Gate,

Reading RG6 6BB, UK

1. Peel, M. C. & McMahon, T. A. *Nature* **444**, doi:10.1038/nature05480 (2006).
2. Labat, D., Godd  ris, Y., Probst, J. L. & Guyot, J. L. *Adv. Water Resour.* **27**, 631–642 (2004).
3. Gedney, N. *et al. Nature* **439**, 835–838 (2006).
4. Legates, D. R., Lins, H. F. & McCabe, G. J. *Adv. Water Resour.* **28**, 1310–1315 (2005).
5. Shiklomanov, I. A. *Water Int.* **25**, 11–31 (2000).
6. Gornitz, V., Rosenzweig, C. & Hillel, D. *Global Planet. Change* **14**, 147–161 (1997).

doi:10.1038/nature05481

PROGRESS

The critical role of disks in the formation of high-mass stars

Riccardo Cesaroni¹, Daniele Galli¹, Giuseppe Lodato², Malcolm Walmsley¹ & Qizhou Zhang³

Although massive stars (commonly defined as those in excess of about eight solar masses, or with initial luminosities of a thousand times the solar luminosity or more) have an enormous impact on the galactic environment, how they form has been a mystery. The solution probably involves the existence of accretion disks. Rotational motions have been found in the gas surrounding young high-mass stars, which suggests that non-spherical accretion could be the fundamental ingredient of the massive-star formation recipe.

High-mass, massive and luminous are all terms used to indicate O-B type stars in excess of ~ 8 solar masses ($8M_{\odot}$). Throughout their lives such stars have a great impact on the interstellar gas, first by ionizing their surroundings by way of their intense ultraviolet radiation, then by exploding as supernovae, which provide chemical enrichment of the galactic medium. But the most intriguing question about massive stars is 'how do they form?'. One cannot simply extrapolate from the case of low-mass stars, where star formation is thought to proceed through accretion onto a dense 'seed'—the protostar¹. Whereas for small protostellar masses the contraction of the nucleus due to self-gravitation is much slower than accretion onto it, when the mass becomes greater than $\sim 8M_{\odot}$ the opposite occurs²: this causes rapid contraction of the protostar, which heats up the stellar interior and ignites hydrogen burning while the star is still accreting.

Theory³ suggests that the pressure of stellar radiation on the infalling material could halt the infall, terminating the accretion process at a mass of $\sim 8M_{\odot}$. However, recent work indicates that this is not a fundamental problem, as discussed below. Here, we suggest that circumstellar accretion disks are likely to play a crucial role in making possible further accretion onto the star. In fact, keplerian circumstellar disks have been detected in a number of low-mass pre-main-sequence stars⁴, which is not surprising as disks seem a natural outcome of the star formation process, owing to conservation of angular momentum during the collapse. In similar fashion, strong pinching forces resulting from the bending of the magnetic field lines trapped in the collapsing cloud produce extended, flattened (albeit non-rotating) density distributions around the accreting protostar⁵. *A priori* there is no reason why these considerations should not hold also for high-mass stars.

The routes to high-mass-star formation

Various models have been proposed to solve the radiation pressure problem. Merging of lower-mass stars in a cluster is an intriguing possibility, although this seems to require stellar densities⁶ greater than 10^6 stars pc^{-3} , well above those observed (for example, 10^4 stars pc^{-3} in the Orion cluster). Alternatively, one may invoke accretion rates much larger than for low-mass stars^{7,8}, thus increasing the ram pressure of the infalling material relative to the radiation pressure of the star. A third possibility is that accretion is not spherically symmetric, as expected for a circumstellar accretion disk. This has a twofold effect: on the one hand, focusing accretion into a small solid

angle boosts the ram pressure of the infalling gas; on the other, lowering the density along the disk axis allows a large fraction of the stellar photons to escape, thus significantly reducing the radiation pressure^{9–14}. This scenario may also involve the presence of a bipolar jet/outflow along the disk axis, possibly driven by centrifugal forces along magnetic field lines rooted in the disk^{15,16}: the result is that part of the infalling material is returned to the surrounding medium. Such outflows are observed, and are indirect evidence for the presence of disks.

Theories of accretion disks

The large infall rates expected in O-B star forming regions also have implications for the structure and stability of the putative associated disks, mostly because the disk masses are expected to be comparable to those of the central stars. Consequently, a theory of disks around O-B stars must address two questions: first, are disks stable against gravitational perturbations? And second, is the transfer of material through the disk sufficient to support accretion rates $\geq 10^{-4} M_{\odot} \text{ yr}^{-1}$ onto the star?

The stability of massive disks requires that gravitational forces be balanced by pressure and centrifugal forces¹⁷. Disks more massive than a few tenths of the mass of their host stars are expected to be unstable. Gravitationally unstable disks might help massive-star formation by providing an efficient way of redistributing angular momentum and promoting accretion through the development of spiral structure¹⁸ (or a bar). However, another possible outcome of the instability is the fragmentation of the disk.

Major advances in this field have been achieved through the use of hydrodynamical simulations. It has been realized^{18,19} that the disk thermodynamics is an essential ingredient in determining both the effectiveness of spiral-structure-induced accretion and the likelihood of fragmentation. A simultaneous treatment of the disk hydrodynamics and radiation physics is just becoming computationally feasible²⁰. Using a simple prescription for the cooling rate, it has been shown that gravitational instabilities saturate at an amplitude at which the dissipation provided by the instability balances the cooling. For larger coolings, the response of the disk is to fragment^{19,21}. Current models show that this sets an upper limit of $10^{-5} M_{\odot} \text{ yr}^{-1}$ to the rate at which accretion can proceed steadily²²; this is about an order of magnitude less than the value needed for massive-star formation, and two orders of magnitude below the mass loss rate measured in the associated bipolar outflows²³ (see above). (It should

¹INAF—Osservatorio Astrofisico di Arcetri, Largo E. Fermi 5, 50125 Firenze, Italy. ²Institute of Astronomy, Cambridge University, Madingley Road, Cambridge CB3 0HA, UK. ³Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, Massachusetts 02138, USA.

be noted that all these results do not include the possibly stabilizing effect of irradiation from the central source.) One can think of a number of possible solutions to this apparent conflict. In particular, it should be remembered that the mass and energy transfer in these disks might be dominated by wave transport, rather than by viscous diffusion. Such non-local effects could significantly affect the accretion rates, although their importance is still controversial^{18,19}.

In summary, some progress has been made, but numerical simulations do not yet fully support a mass transfer through the disk that is sufficiently high to 'feed' the (proto)star at rates of 10^{-4} – $10^{-3} M_{\odot} \text{ yr}^{-1}$.

Searching for disks

A number of searches for circumstellar disks have been performed towards selected targets. These are usually luminous infrared sources (mostly selected from the IRAS point source catalogue), maser emitters, very compact (that is, very young) ionized regions, as well as hot (~ 100 K), small (< 0.1 pc) molecular condensations named 'hot cores'²⁴. Detection of a bipolar molecular outflow is an additional selection criterion, given their tight association with disks (see above). Outflows are much easier to detect, being much larger than disks and characterized by broad wings in the mm lines of CO. Surveys searching for outflows towards high-mass young stellar objects (YSOs) report detection rates^{25–28} of up to 90%, suggesting that disks are common in these sources.

Disk searches require the choice of a good tracer: several points need to be considered. Disks are expected to be deeply embedded in the innermost, densest parts of molecular cores. Distinguishing the disk emission from that of the surrounding envelope requires an optically thin tracer, such as (sub)mm continuum emission from the dust or line emission from rare molecular species that trace the dense gas in disks. For these reasons, most observations are done at (sub)mm wavelengths, where many molecular lines are emitted and the dust continuum emission is optically thin and not contaminated by free-free emission from possibly associated ionized gas. Alternatively, the (optically thick) disk can be seen in silhouette against a bright background in the near-infrared^{29–31} or in emission in the mid-infrared³¹. In any case, high sensitivity and angular resolution are needed, because the expected disk diameter is $< 1,000$ AU, which corresponds to $< 1''$ for distances > 1 kpc, typical of high-mass-star forming regions. Interferometric observations at (sub)mm wavelengths and imaging in the mid/near-infrared fulfil such requirements.

A special case is that of maser emission. Several molecular species in high-mass-star forming regions have been found to undergo population inversion in some of their transitions: this causes amplification of the corresponding lines by stimulated emission of radiation, comparable to laboratory masers. Observationally, one detects strong lines (up to 10^{10} K in brightness temperature) arising from very small (~ 1 AU) regions, named 'spots'. These are ideal targets for observations with very long baseline interferometry (VLBI), with resolutions of the order of 1 mas. This technique has been used to sample the structure and velocity field of the gas close to embedded, massive stellar objects and thus possibly infer the presence of circumstellar disks³² and/or jets.

Advantages and shortcomings of continuum, thermal lines and maser lines for disk searches are summarized in Table 1. Although continuum observations may benefit from larger bandwidths and thus from better sensitivity, only line observations allow studies of the gas kinematics. In particular, proper motion and radial velocity measurements of the maser spots make it possible to reconstruct the

three-dimensional velocity field. Investigation of the kinematics plays a crucial role in disk identification. Disentangling the disk from the surrounding envelope is difficult using only density and temperature information, as the boundary between the two may not be easily identifiable. The best way to infer the presence of a disk is to establish the existence of rotation. However, the detection of a velocity gradient along a well defined direction does not by itself guarantee that one is observing rotation, as other types of motion (such as expansion or infall) can produce the same effect. The association between disks and jets/outflows provides us with one method to discriminate between different interpretations. As the outflow direction is close to the disk axis, one may compare the direction of the velocity gradient detected in a high-density tracer on a small scale (~ 0.1 pc) with that of the larger-scale (~ 1 pc) outflow: if the two are nearly perpendicular to one another, it is plausible that the small-scale velocity gradient corresponds to rotation about the outflow axis. The power of this technique is enhanced by the fact that the large bandwidths covered by the receivers and spectrum analysers of interferometers operating at (sub)mm wavelengths allow simultaneous measurements of disk and outflow tracers.

Results of disk searches

In recent years, a number of candidate disks in high-mass YSOs have been reported in the literature^{29–43}. These may be roughly classified into two classes on the basis of their luminosities and stellar masses, the dividing line being set at $\sim 10^5 L_{\odot}$ or $\sim 25 M_{\odot}$ for a main-sequence star.

YSOs above this limit are characterized by the presence of large (~ 0.1 pc), massive (typically a few hundred $M_{\odot}$) rotating structures (see ref. 22 and references therein), whose rotation period ($\sim 10^5$ yr) is longer than the time required to accrete the whole rotating mass onto a central star, at accretion rates of $\sim 10^{-3}$ – $10^{-2} M_{\odot} \text{ yr}^{-1}$ (ref. 44). This suggests that such structures never reach equilibrium and are continuously re-created through infall of material from the parental cloud. Because of this and the large masses involved, it is likely that these rotating structures are not enshrouding a single star but rather a whole cluster. To distinguish them from the 'circumstellar disks' seen in YSOs below $\sim 10^5 L_{\odot}$, the term 'circumcluster toroids' has been adopted. Ordinary disks, on the other hand, are characterized by masses less than or comparable to that of the central (proto)star (4 – $20 M_{\odot}$) and accretion rates of $\sim 10^{-4} M_{\odot} \text{ yr}^{-1}$, which in turn imply accretion timescales longer than the rotation period ($\sim 10^4$ yr). As previously discussed, structures like these may be stable and relatively long-lived ($\sim 10^5$ yr). Unlike the massive toroids, in this case the mass of the star dominates the gravitational potential and keplerian or quasi-keplerian⁴⁵ rotation is expected in the outer region of the disk, while at smaller radii accretion is likely to dominate and rotation is expected to turn into infall onto the star.

Evidence of circumstellar disks has been found in a dozen objects. These have masses of 0.2 – $40 M_{\odot}$ and radii of 500 – $10,000$ AU, and are associated with bipolar molecular outflows with mass loss rates of 10^{-5} – $10^{-3} M_{\odot} \text{ yr}^{-1}$ and dynamical timescales of a few times 10^4 yr. The most representative—and best studied—is IRAS 20126+4104, a luminous ($\sim 10^4 L_{\odot}$) far-infrared source associated with a bipolar jet/outflow³⁸. The geometry is illustrated in Fig. 1. The protostar, located at the centre of the flow, is embedded in a dense, compact, hot molecular core, which is found to undergo quasi-keplerian rotation about a $\sim 7 M_{\odot}$ protostar³⁸. VLBI measurements of the proper motions of the water maser spots⁴⁶ have revealed expansion at $\sim 100 \text{ km s}^{-1}$ along the jet axis, on scales as small as a few tens of

Table 1 | Tracers used for disk searches and their characteristics

Tracer	Advantages	Shortcomings
Continuum millimetre and sub-millimetre Thermal lines of CH_3CN , NH_3 and so on Maser lines of H_2O , CH_3OH	Large bandwidth gives sensitivity Gives kinematics and geometry of outflow and disk Very high (1 mas) resolution proper motions give three-dimensional velocity	No velocity information; confusion with free-free and envelope Limited ($\sim 1''$) angular resolution and sensitivity Sample only limited parts of small (< 100 AU) regions

AU. The conclusion is that one is observing a nice example of a massive (proto)star associated with a disk-outflow system.

An important finding is that all YSOs in disk candidates have luminosities typical of B-type main-sequence stars, that is, $<10^5 L_{\odot}$: to date, no circumstellar disk has been found around early O-type stars, but only massive rotating toroids. Radio source I in Orion might be an exception, but in this case the luminosity estimate is highly uncertain (between 4×10^3 and $10^5 L_{\odot}$) and recent measurements of the free-free continuum emission⁴⁷ are consistent with a spectral type B1 for the star at the centre of the disk.

Given the large distances to O stars, it is possible that the lack of evidence for disks in O-type (proto)stars is due to observational bias. However, other explanations are possible. On the one hand, the circumstellar disks in O stars might be 'hidden' inside the toroids and hence very difficult to disentangle from the surrounding larger-scale rotating structure. In this way, deeply embedded, heavily accreting O protostars might 'mimic' B-type stars surrounded by massive disks⁴⁸. Only interferometers such as the Atacama Large Millimeter Array (ALMA) will provide the angular resolution and sensitivity needed to shed light on this issue. On the other hand, disks in O stars could be rapidly destroyed or never created. The latter hypothesis implies a different formation mechanism for O stars, as opposed to lower-mass stars. Theories involving merging of lower-mass stars⁶ have been proposed, as well as models where the cluster plays a crucial role, deepening the gravitational well around the most massive star and boosting accretion onto it⁸. In both cases, the circumstellar environment is strongly perturbed and the disk could be severely affected by interactions with stellar companions. This brings us to the question of disk life time, which is not only influenced by tidal interactions with the cluster, but also by the powerful ionizing flux of the early-type star.

The effect of the ionizing stellar radiation from the central star on the disk has been modelled in detail^{49,50}, with the conclusion that the mass loss rate due to photoevaporation is 100–1,000 times less than the accretion rate onto the star. However, a complete model (that takes into account the interplay between transfer of material in the disk and photo-evaporation) is needed before drawing any conclusion, especially in the case of O stars, the most powerful Lyman continuum emitters.

The other type of influence on disk lifetime—tidal interactions with stellar companions in binary (multiple) systems and with cluster members—depends on the distance of the disk from the companion(s) and on the stellar density. A simplified approach^{22,49} demonstrates that the typical timescale for destroying the outer regions of a 1,000 AU disk is $\sim 10^5$ yr, whereas tidal effects due to the associated cluster may affect the disk if the radius is $>3,500$ AU and the stellar density exceeds $\sim 4 \times 10^4$ stars pc⁻³. In conclusion, it is possible that truncation of circumstellar disks around massive stars may occur, which may have an important impact on their observability. This might explain the previously mentioned apparent lack of disks in O stars.

Future prospects

Despite some recent success in the search for circumstellar accretion disks in early-type YSOs, robust evidence for their existence has been found only in B-type (proto)stars, whereas disks in O stars appear to be elusive. This is due to observational bias or to disk truncation caused by interactions with the associated stellar cluster. However, the most extreme possibility—that O stars form through a completely different mechanism, not involving accretion—cannot be excluded. The superb sensitivity and resolution attained by new generation instruments such as ALMA will certainly allow substantial progress towards a complete understanding of the high-mass-star formation process. Significant advance is also expected on the theoretical side, with an improvement in numerical modelling taking into account the disk physics in three dimensions and its interaction with the star and associated cluster. A new era for high-mass-star formation studies is ahead of us.

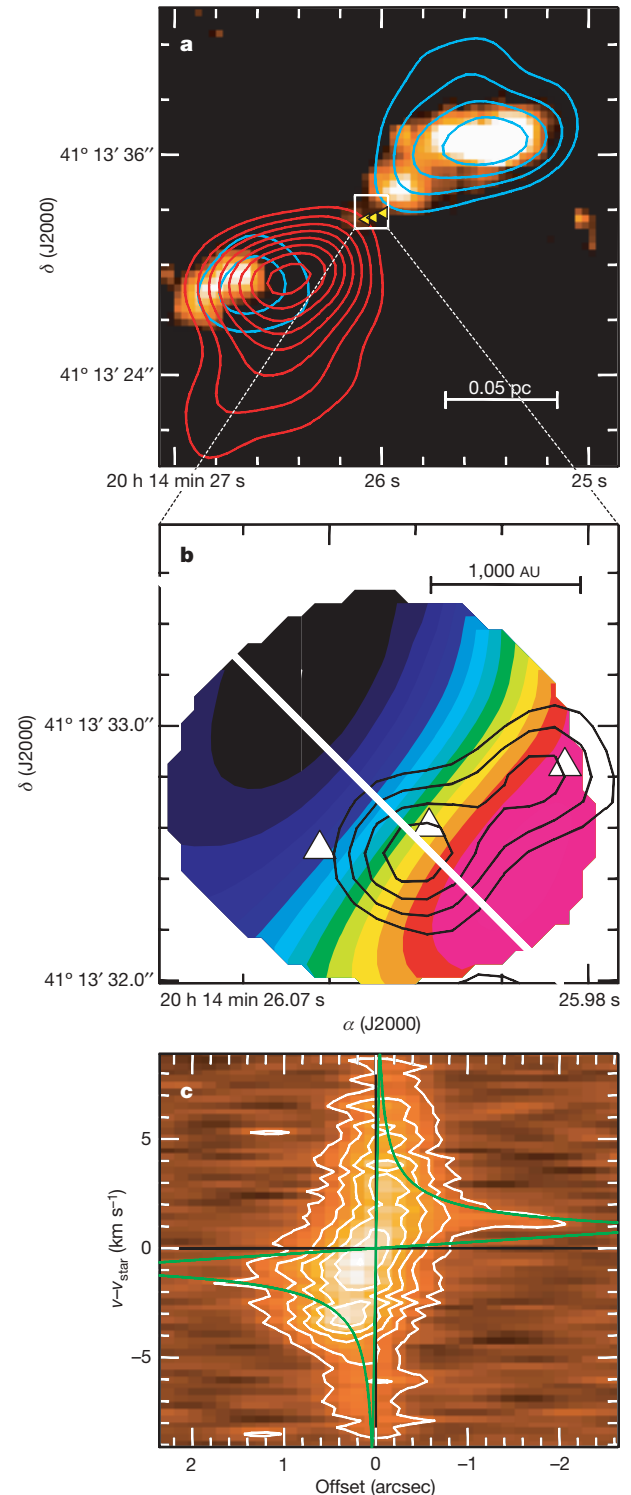


Figure 1 | Illustration of the circumstellar disk and associated bipolar outflow in the high-mass (proto)star IRAS 20126+4104. **a**, Overlay of the H₂ line emission at 2.2 μm (colour scale) and the bipolar outflow traced by the HCO⁺ (1–0) line. Blue and red contours correspond respectively to blue- and red-shifted gas. The triangles mark the positions of the H₂O maser spots. **b**, 3.6 cm continuum map (contours) overlaid on a map of the velocity measured in the C³⁴S(5–4) line (colour scale ranging from red for red-shifted to blue for blue-shifted emission). Panel shows a magnified view of the boxed area in **a**. White triangles mark the positions of water maser spots. **c**, Intensity map (contours and colour scale), where white corresponds to the maximum intensity) of the C³⁴S(5–4) line emission as a function of velocity relative to the star ($v - v_{\text{star}}$) and positional offset along the plane of the disk, outlined by the white line in **b**. The green pattern denotes the region from which emission is expected in the case of keplerian rotation about a $7M_{\odot}$ star. The black lines correspond to zero velocity and offset.

1. Shu, F. H., Adams, F. C. & Lizano, S. Star formation in molecular clouds—Observation and theory. *Annu. Rev. Astron. Astrophys.* **25**, 23–81 (1987).
2. Palla, F. & Stahler, S. W. The pre-main-sequence evolution of intermediate-mass stars. *Astrophys. J.* **418**, 414–425 (1993).
3. Kahn, F. D. Cocoons around early-type stars. *Astron. Astrophys.* **37**, 149–162 (1974).
4. Simon, M., Dutrey, A. & Guilloteau, S. Dynamical masses of T Tauri stars and calibration of pre-main-sequence evolution. *Astrophys. J.* **545**, 1034–1043 (2000).
5. Galli, D. & Shu, F. H. Collapse of magnetized molecular cloud cores. II. Numerical results. *Astrophys. J.* **417**, 243–258 (1993).
6. Bonnell, I. A. & Bate, M. R. Binary systems and stellar mergers in massive star formation. *Mon. Not. R. Astron. Soc.* **362**, 915–920 (2005).
7. McKee, C. F. & Tan, J. C. Massive star formation in 100,000 years from turbulent and pressurized molecular clouds. *Nature* **416**, 59–61 (2002).
8. Bonnell, I. A., Vine, S. G. & Bate, M. R. Massive star formation: nurture, not nature. *Mon. Not. R. Astron. Soc.* **349**, 735–741 (2004).
9. Nakano, T. Conditions for the formation of massive stars through nonspherical accretion. *Astrophys. J.* **345**, 464–471 (1989).
10. Jijina, J. & Adams, F. C. Infall collapse solutions in the inner limit: Radiation pressure and its effects on star formation. *Astrophys. J.* **462**, 874–887 (1996).
11. Yorke, H. W. & Sonnhalter, C. On the formation of massive stars. *Astrophys. J.* **569**, 846–862 (2002).
12. Krumholz, M. R., McKee, C. F. & Klein, R. I. The formation of stars by gravitational collapse rather than competitive accretion. *Nature* **438**, 332–334 (2005).
13. Krumholz, M. R., Klein, R. I. & McKee, C. F. in *Massive Star Birth: A Crossroads of Astrophysics* (eds Cesaroni, R., Felli, M., Churchwell, E. & Walmsley, M.) 231–236 (IAU Symp. 227, Cambridge Univ. Press, Cambridge, UK, 2005).
14. Krumholz, M. R., McKee, C. F. & Klein, R. I. How protostellar outflows help massive stars form. *Astrophys. J.* **618**, L33–L36 (2005).
15. Shu, F. H., Najita, J. R., Shang, H. & Li, Z.-Y. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 789–813 (Univ. Arizona Press, Tucson, 2000).
16. Ferreira, J. & Pelletier, G. Magnetized accretion-ejection structures. III. Stellar and extragalactic jets as weakly dissipative disk outflows. *Astron. Astrophys.* **295**, 807–832 (1995).
17. Toomre, A. On the gravitational stability of a disk of stars. *Astrophys. J.* **139**, 1217–1238 (1964).
18. Lodato, G. & Rice, W. K. M. Testing the locality of transport in self-gravitating accretion discs – II. The massive disc case. *Mon. Not. R. Astron. Soc.* **358**, 1489–1500 (2005).
19. Mejia, A. C., Durisen, R. H., Pickett, M. K. & Cai, K. The thermal regulation of gravitational instabilities in protoplanetary disks. II. Extended simulations with varied cooling rates. *Astrophys. J.* **619**, 1098–1113 (2005).
20. Whitehouse, S. C. & Bate, M. R. Smoothed particle hydrodynamics with radiative transfer in the flux-limited diffusion approximation. *Mon. Not. R. Astron. Soc.* **353**, 1078–1094 (2004).
21. Rice, W. K. M., Lodato, G. & Armitage, P. J. Investigating fragmentation conditions in self-gravitating accretion discs. *Mon. Not. R. Astron. Soc.* **364**, L56–L60 (2005).
22. Cesaroni, R., Galli, D., Lodato, G., Walmsley, C. M. & Zhang, Q. in *Protostars and Planets V* (eds Reipurth, B., Jewitt, D. & Keil, K.) (Univ. Arizona Press, Tucson, in the press).
23. Shepherd, D. in *Massive Star Birth: A Crossroads of Astrophysics* (eds Cesaroni, R., Felli, M., Churchwell, E. & Walmsley, M.) 237–246 (IAU Symp. 227, Cambridge Univ. Press, Cambridge, UK, 2005).
24. Kurtz, S., Cesaroni, R., Churchwell, E., Hofner, P. & Walmsley, C. M. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 299–326 (Univ. Arizona Press, Tucson, 2000).
25. Shepherd, D. S. & Churchwell, E. High-velocity molecular gas from high-mass star formation regions. *Astrophys. J.* **457**, 267–276 (1996).
26. Osterloh, M., Henning, Th. & Launhardt, R. Infrared images and millimeter data from cold southern IRAS sources. *Astrophys. J.* **110** (Suppl.), 71–114 (1997).
27. Zhang, Q. et al. Search for CO outflows toward a sample of 69 high-mass protostellar candidates: Frequency of occurrence. *Astrophys. J.* **552**, L167–L170 (2001).
28. Beuther, H. et al. Massive molecular outflows. *Astron. Astrophys.* **383**, 892–904 (2002).
29. Chini, R. et al. The formation of a massive protostar through the disk accretion of gas. *Nature* **429**, 155–157 (2004).
30. Sako, S. et al. No high-mass protostars in the silhouette young stellar object M17–SO1. *Nature* **434**, 995–998 (2005).
31. Sridharan, T. K., Williams, S. J. & Fuller, G. A. The direct detection of a (proto)binary/disk system in IRAS 20126+4104. *Astrophys. J.* **631**, L73–L76 (2005).
32. Greenhill, L. J., Reid, M. J., Chandler, C. J., Diamond, P. J. & Elitzur, M. in *Star Formation at High Angular Resolution* (eds Burton, M., Jayawardhana, R. & Bourke, T.) 155–160 (IAU Symp. 221, Kluwer/Springer, Dordrecht, 2004).
33. Schreyer, K., Henning, Th., van der Tak, F. F. S., Boonman, A. M. S. & van Dishoeck, E. F. The young intermediate-mass stellar object AFGL 490 – A disk surrounded by a cold envelope. *Astron. Astrophys.* **394**, 561–583 (2002).
34. Shepherd, D. S., Claussen, M. J. & Kurtz, S. E. Evidence for a solar system-size accretion disk around the massive protostar G192.16–3.82. *Science* **292**, 1513–1518 (2001).
35. Zhang, Q., Hunter, T. R., Sridharan, T. K. & Ho, P. T. P. A disk/jet system toward the high-mass young star in AFGL 5142. *Astrophys. J.* **566**, 982–992 (2002).
36. Bernard, J. P., Dobashi, K. & Momose, M. Outflow and disk around the very young massive star GH20 092.67+03.07. *Astron. Astrophys.* **350**, 197–203 (1999).
37. Jiang, Z. et al. A circumstellar disk associated with a massive protostellar object. *Nature* **437**, 112–115 (2005).
38. Cesaroni, R. et al. A study of the Keplerian accretion disk and precessing outflow in the massive protostar IRAS 20126+4104. *Astron. Astrophys.* **434**, 1039–1054 (2005).
39. Hutawarakorn, B. & Cohen, R. J. Magnetic field structure in the bipolar outflow source G 35.2–0.74N: MERLIN spectral line results. *Mon. Not. R. Astron. Soc.* **303**, 845–854 (1999).
40. Patel, N. A. et al. A disk of dust and molecular gas around a high-mass protostar. *Nature* **437**, 109–111 (2005).
41. van der Tak, F. F. S. & Menten, K. M. Very compact radio emission from high-mass protostars. II. Dust disks and ionized accretion flows. *Astron. Astrophys.* **437**, 947–956 (2005).
42. Beuther, H., Zhang, Q., Sridharan, T. K. & Chen, Y. Testing the massive disk scenario for IRAS 18089–1732. *Astrophys. J.* **628**, 800–810 (2005).
43. Chini, R. et al. A remnant disk around a young massive star. *Astrophys. J.* **645**, L61–L64 (2006).
44. Zhang, Q. in *Massive Star Birth: A Crossroads of Astrophysics* (eds Cesaroni, R., Felli, M., Churchwell, E. & Walmsley, M.) 135–144 (IAU Symp. 227, Cambridge Univ. Press, Cambridge, UK, 2005).
45. Bertin, G. & Lodato, G. A class of self-gravitating accretion disks. *Astron. Astrophys.* **350**, 694–704 (1999).
46. Moscadelli, L., Cesaroni, R. & Rioja, M. J. Water masers in the massive protostar IRAS 20126+4104: ejection and deceleration. *Astron. Astrophys.* **438**, 889–898 (2005).
47. Beuther, H. et al. Submillimeter array 440 μ m/690 GHz line and continuum observations of Orion KL. *Astrophys. J.* **636**, 323–331 (2006).
48. Beuther, H. & Shepherd, D. in *Cores to Clusters* (eds Kumar, M. S. N., Tafalla, M. & Caselli, P.) 105–119 (Springer, New York, 2005).
49. Hollenbach, D. J., Yorke, H. W. & Johnstone, D. in *Protostars and Planets IV* (eds Mannings, V., Boss, A. P. & Russell, S. S.) 401–428 (Univ. Arizona Press, Tucson, 2000).
50. Richling, S. & Yorke, H. W. Photoevaporation of protostellar disks. II. The importance of UV dust properties and ionizing flux. *Astron. Astrophys.* **327**, 317–324 (1997).

Acknowledgements We thank M. Felli for a critical reading of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.C. (cesa@arcetri.inaf.it).

Electrical activity in early neuronal development

Nicholas C. Spitzer¹

The construction of the brain during embryonic development was thought to be largely independent of its electrical activity. In this view, proliferation, migration and differentiation of neurons are driven entirely by genetic programs and activity is important only at later stages in refinement of connections. However, recent findings demonstrate that activity plays essential roles in early development of the nervous system. Activity has similar roles in the incorporation of newly born neurons in the adult nervous system, suggesting that there are general rules underlying activity-dependent development. The extensive involvement of activity makes it likely that it is required at all developmental stages as a necessary partner with genetic programs.

Early neuronal development may be divided into three major phases occurring sequentially but partially overlapping. Proliferation from stem cells generates large numbers of neurons. Migration of neurons to achieve their final locations follows on the heels of proliferation. Differentiation includes maturation of electrical excitability, transmitter specification, and outgrowth of axons and dendrites. Here I consider newly discovered functions of electrical activity in these events that follow neural induction and precede synapse formation.

The activity dependence of synapse formation and neuronal survival at later stages of development has been recognized for some time and has been recently reviewed^{1–4}. Proliferation, migration and differentiation all depend to various extents on forms of electrical activity, both embryonically and in the adult. Given the involvement of ion channel signalling in neural induction⁵, it appears that electrical activity plays a role at all stages of development. Research now focuses on the mechanisms by which this activity exerts its effects.

Conventional forms of excitability

In the past, electrical activity was viewed exclusively as the activity of action potentials in which the inward current is carried by Na ions. Because tetrodotoxin blocks voltage-gated Na channels and the action potentials they generate, tetrodotoxin was the principal assay used to determine whether or not a developmental process is activity-dependent. Recognition of the role of Ca as a second messenger coupled with the development of techniques for imaging intracellular Ca ion concentrations led to the discovery of spontaneous fluctuations in intracellular Ca at early stages of neuronal development (Box 1). Examination of the origins of these fluctuations revealed roles for neurotransmitters that depolarize neurons and activate voltage-gated Ca channels that promote entry of Ca, or that bind to receptors that flux Ca. The transmitters GABA (γ -aminobutyric acid) and glycine activate receptors that generate depolarizing chloride currents at early stages, while glutamate receptors generate Na and Ca currents that depolarize neurons throughout development. As a result, tests of the roles of these forms of excitability in neuronal development require blockade not only of voltage-gated Na channels, but voltage-gated Ca channels and neurotransmitter receptors, for both of which there are large families. Proliferation, migration, ion channel

expression, neurotransmitter specification, axon pathfinding and dendrite outgrowth are all regulated by the expanded repertoire of conventional forms of excitability.

Novel forms of excitability

The discovery of previously unknown forms of excitability has also contributed to the appreciation of the role of electrical activity in neuronal development. Here excitability is defined as the influx of ions, often Ca ions, via mechanisms that do not depend on classical voltage-gated or neurotransmitter-gated channels. For example, TRP (transient receptor potential) channels are involved in the generation of growth cone Ca transients. Metabotropic transmitter receptors, mechanoreceptors, and transmitter transporters acting in reverse are less-well-studied components generating electrical signals that can shape neuronal development. Tests for the roles of Ca transients in

Box 1 | Spontaneous activity before synapse formation

Observations of spontaneous electrical activity at early stages of development before synapse formation suggested that it could have roles in the assembly of the nervous system. Electrical recordings and optical imaging of intracellular Ca revealed the presence of extensive neurotransmitter-evoked activity in a range of CNS structures in both vertebrates and invertebrates before the appearance of conventional synaptic structures. Endogenous release of a neurotransmitter activates NMDA receptors in embryonic turtle cortex⁷² and generates Ca transients via activation of metabotropic glutamate receptors in embryonic mouse cortex⁷³. Endogenous transmitters also engage GABA_A receptors that lead to depolarization of neuronal progenitors in the postnatal mouse subventricular zone and rostral migratory stream⁷⁴.

Spontaneous Ca transients are also triggered by substrate interactions and by mechanisms that are still unknown. Ca transients are observed in embryonic mouse neural crest⁷⁵, *Xenopus* and zebrafish spinal neurons^{20,32,42,76–78}, chick dorsal root ganglion neurons³¹ and hamster cortical neurons³³. Several different patterns of Ca transients are observed in precursor cells of the embryonic rat neocortical ventricular zone⁷⁹. In tunicates and invertebrate nervous systems, spontaneous Ca elevations are generated in cells as varied as embryonic ascidian muscle and pupal moth antennal-lobe and fly mushroom-body neurons^{19,80,81}.

¹Neurobiology Section, Division of Biological Sciences and Centre for Molecular Genetics, Kavli Institute for Brain and Mind, University of California San Diego, La Jolla, California 92093-0357, USA.

developmental processes now include imaging of intracellular Ca during the relevant period of development to learn whether such forms of excitability are present. It seems likely that further classes of Ca transients await identification.

Neuronal proliferation

Neurons are generated in vast numbers in the embryonic brain, as asymmetric divisions of neural progenitors give rise to neurons at a prodigious rate. Strikingly, this process is regulated by electrical signalling. GABA and glutamate released at this early stage before synapse formation have different effects on neurogenesis in different regions. GABA and glutamate depolarize embryonic rat cortical progenitor cells in the ventricular zone, stimulate elevations of intracellular Ca, and inhibit DNA synthesis. Blockade of GABA_A receptors and the AMPA (α -amino-3-hydroxy-5-methyl-4-isoxazole propionic acid)/kainate class of glutamate receptors increases DNA synthesis, indicating a role for endogenously released transmitters in regulating proliferation⁶. Similarly, release of GABA from differentiating neuroblasts depolarizes their progenitor cells in the postnatal mouse subventricular zone and limits proliferation⁷ (Fig. 1). In contrast, GABA induces proliferation of postnatal rat immature cerebellar granule cells by depolarization and activation of Ca channels⁸. Disruption of Ca wave propagation through embryonic rat radial glia that are neuronal progenitors decreases neuronal proliferation in the ventricular zone⁹. Roles for other neurotransmitters and the mechanisms by which activity stimulates or inhibits proliferation are now subjects of interest. The sign of the effects of electrical activity can be regulated by the ratio of 3', 5'-cyclic AMP to cyclic GMP^{10,11}.

Neuronal migration

Neurons are generated in a restricted set of locations and generally migrate to specific sites from which they extend processes to establish synaptic connections. Some neurons migrate radially and others migrate tangentially. Both the rate and extent of migration are regulated by neurotransmitters released at this early point in development. Radial migration of postnatal mouse cerebellar granule cells along Bergmann glia depends on activation of voltage-gated N-type Ca channels and the NMDA (N-methyl-D-aspartate) class of glutamate receptors, which generate fluctuations in intracellular Ca ions that are positively correlated with their rate of migration¹². Release of GABA and glutamate in a manner that does not require conventional SNARE-dependent exocytosis of vesicle contents promotes migration of embryonic mouse hippocampal pyramidal neurons^{13,14}. Somatostatin increases the frequency of Ca transients and the initial rate of postnatal murine granule cell migration as well as the loss of Ca transients that triggers the completion of migration¹⁵ (Fig. 2). In contrast, GABA slows the rate of migration in other regions of the brain. Astrocyte-like cells regulate the rate of tangential

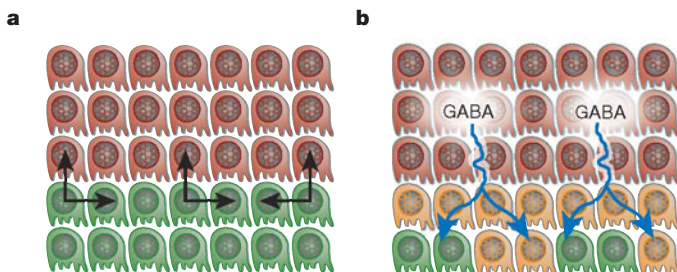


Figure 1 | Spontaneous release of neurotransmitters from neuroblasts generates elevations of calcium in embryonic progenitor cells that suppress their proliferation. **a**, Cerebral cortical progenitor cells (green) in the ventricular zone divide asymmetrically (black arrows) to give rise to neuroblasts (red) in rodent brains. **b**, Neuroblasts release GABA (blue arrows), depolarizing progenitors and stimulating elevations of Ca (yellow) that inhibit DNA synthesis and proliferation. Adapted from ref. 7, with permission.

migration by uptake of GABA in the postnatal mouse subventricular zone and rostral migratory stream; migration is reduced by inhibiting uptake and increased by blocking GABA_A receptors¹⁶. The signal transduction pathways by which neurotransmitters stimulate or inhibit neuronal migration are now being investigated^{17,18}; the ratio of cAMP to cGMP appears to regulate the sign of this process as well.

Neuronal differentiation

Voltage-gated ion channel expression. Neuronal differentiation has many components, one of which is the specification of ion channels that govern excitability. Spontaneous electrical activity can regulate the level of excitability of cells by controlling the expression of ion channels generating it. The best-studied example is the development of embryonic ascidian muscle, in which Ca currents are expressed before the appearance of rapidly activated outward and inwardly rectifying potassium currents, engendering a period of spontaneous activity. When Ca-dependent action potentials are blocked, developmental expression of the rapidly activated potassium current is specifically suppressed¹⁹. In embryonic *Xenopus* spinal neurons the delayed rectifier current plays the central role in developmental conversion of Ca-dependent action potentials to Na-dependent spikes. Ca influx through voltage-gated channels at early stages of development drives an increase in the rate of activation of this current^{20,21}. These results demonstrate feedback loops by which early forms of electrical activity regulate intrinsic excitability and can limit the embryonic period during which activity triggers developmental programs.

Neurotransmitter specification. A second feature of neuronal differentiation is the specification of neurotransmitter phenotype. This process depends on early electrical activity in a number of systems, in conjunction with cell-specific transcription factor expression or action of trophic factors. Alterations in activity can change the numbers of neurons expressing excitatory and inhibitory transmitters. Different classes of embryonic *Xenopus* spinal neurons generate

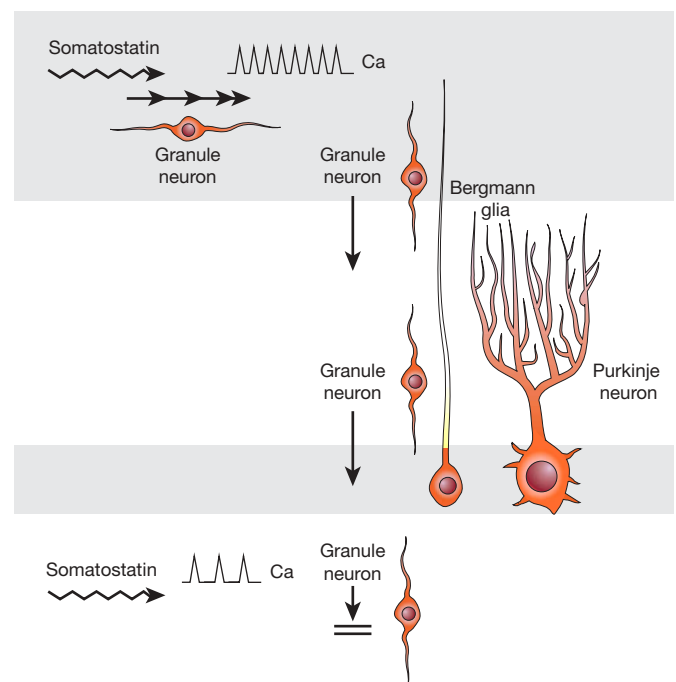


Figure 2 | Neurotransmitter modulation of embryonic neuronal migration by stage-specific regulation of calcium transients. Spontaneous release of somatostatin increases the frequency of Ca transients (Ca) and enhances tangential migration of granule neurons on the superficial layer of the mouse cerebellum. Somatostatin suppresses Ca transients and slows radial migration of granule neurons in the deep layer at a later stage of development. Adapted from ref. 15, with permission.

spontaneous Ca spikes at different frequencies, before synapse formation. In a homeostatic manner, spike suppression increases the incidence of excitatory transmitter expression and decreases the incidence of inhibitory transmitters. Conversely, increasing spike production decreases the incidence of neurons expressing excitatory transmitters and increases the incidence of neurons expressing inhibitory transmitters²² (Fig. 3). Embryonic chick and rat neurons generate developmentally transient Ca-dependent action potentials that may produce Ca spikes like those described for *Xenopus* spinal neurons and regulate transmitter expression similarly^{23–25}. Results from studies of other systems are consistent with activity-dependent homeostatic regulation of transmitter expression. Stimulation of embryonic rat sensory petrosal ganglion neurons *in vitro* induces expression of tyrosine hydroxylase²⁶, a synthetic enzyme for dopamine that acts as an inhibitory transmitter for these neurons. Physiological stimulation of these neurons *in vivo* induces tyrosine hydroxylase expression exclusively in cells expressing *Phox2a/2b* (paired-like homeobox 2a and 2b) transcription factors²⁷, illustrating the partnership between transcription factor expression and activity.

Transmitter receptor populations change to match activity-dependent changes in transmitter expression at the vertebrate neuromuscular junction. Matching of transmitters with their cognate receptors appears to be achieved by a process of selection. Embryonic *Xenopus* muscle cells initially express receptors for glutamate, GABA and glycine as well as acetylcholine; acetylcholine receptor expression prevails as maturation progresses. However the receptor populations remaining in muscle cells parallel changes in transmitter phenotype when neuronal activity is perturbed. Glutamatergic, GABAergic and glycinergic synaptic currents are recorded from muscle cells when early neuronal Ca-dependent activity is manipulated to increase the number of neurons expressing these transmitters⁹⁶. These results are consistent with release of glutamate from neonatal mammalian motoneurons^{28,29} and glutamatergic transmission at adult vertebrate neuromuscular junctions³⁰.

Axon pathfinding. A third aspect of neuronal differentiation is the targeting of axons that neurons extend to synapse with other neurons, muscles or glands. Both novel and conventional forms of activity are involved in pathfinding of axons via the migration of growth cones at their tips. Intracellular Ca has diverse effects on axon extension, turning and branching that may depend on concentration and route of entry. Growth cone motility and neurite extension are inhibited by application of transmitters that elicit depolarizations and Ca increases, and axon extension is inhibited by spontaneous growth

cone Ca transients as well^{20,31–33}. TRPC5-receptor-operated ion channels appear to constitute the Ca entry pathway generating spontaneous growth cone transients in cultured hippocampal neurons³⁴.

Axons turn in response to diffusible guidance molecules. Growth cones of cultured *Xenopus* spinal neurons exhibit Ca-dependent attraction to gradients of acetylcholine and of glutamate³⁵. Ca also mediates growth cone turning of these neurons stimulated by an extracellular gradient of netrin-1 (ref. 36), a diffusible guidance factor *in vivo*. The ratio of cAMP to cGMP modulates the activity of voltage-gated Ca channels that regulate Ca increases and growth cone turning³⁷. The turning responses to netrin-1, brain-derived neurotrophic factor and MAG (myelin-associated glycoprotein) are mediated by TRPC1 channels^{38,39} in *Xenopus* spinal neurons and TRPC3 channels in rat cerebellar granule neurons⁴⁰ (Fig. 4a). Stimulation of action potentials enhances netrin-1 attraction and converts MAG repulsion to attraction, by elevating both Ca and cAMP (ref. 41), demonstrating an intersection of conventional and novel forms of excitability. Axons also turn in response to different substrates. Filopodia that extend from neuronal growth cones sample the environment for extracellular guidance cues, and generate localized transient elevations of intracellular Ca at their tips that propagate back to the growth cone in embryonic *Xenopus* spinal neurons. The frequency of filopodial transients is substrate-dependent and reduces filopodial motility, promoting turning when stimulated differentially within filopodia on one side of the growth cone⁴².

Axons branch both at their tips and interstitially, and both classes of branching are activity-dependent. Axons of retinal ganglion cells arriving in the optic tectum of *Xenopus* tadpoles add branches at their tips equally but eliminate them selectively through a NMDA-receptor-dependent mechanism in regions of the tectum constructed to be dominated by the opposite eye⁴³. Competition between neighbouring axons growing and branching in the zebrafish optic tectum involves activity-dependent neurosecretion that implicates the role of neurotransmitters⁴⁴. Interstitial or collateral branching at points along axons involves chemotropic signals⁴⁵. Netrin-1 increases interstitial branching of cultured hamster somatosensory cortical neurons, stimulating repetitive Ca transients that are spatially and temporally correlated with the sites of axon protrusion; suppression of these transients blocks branching⁴⁶.

Axon guidance by both substrate and diffusible molecules requires expression of appropriate receptors, and receptor expression is activity-dependent in a number of cases^{47,48}. This is an important but relatively unexplored subject.

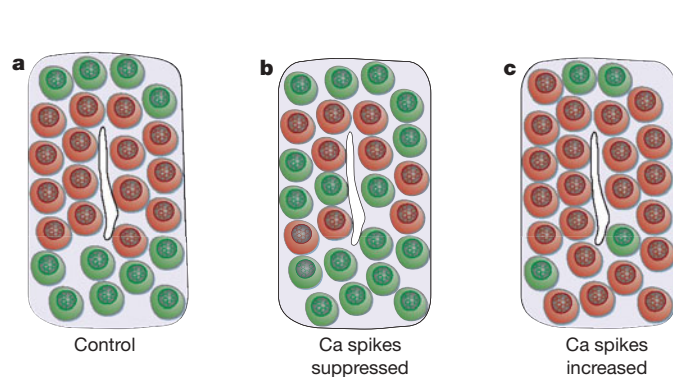


Figure 3 | Calcium-spike-dependent homeostatic specification of embryonic neurotransmitter expression. **a**, Normal expression of excitatory transmitters (green) and inhibitory transmitters (red) in the frog spinal cord. **b**, Suppressing production of spontaneous Ca spikes increases the number of neurons expressing excitatory transmitters and decreases the number of neurons expressing inhibitory transmitters. **c**, Increasing the frequency of Ca spikes decreases the number of neurons expressing excitatory transmitters and increases the number of neurons expressing inhibitory transmitters. Adapted from ref. 22.

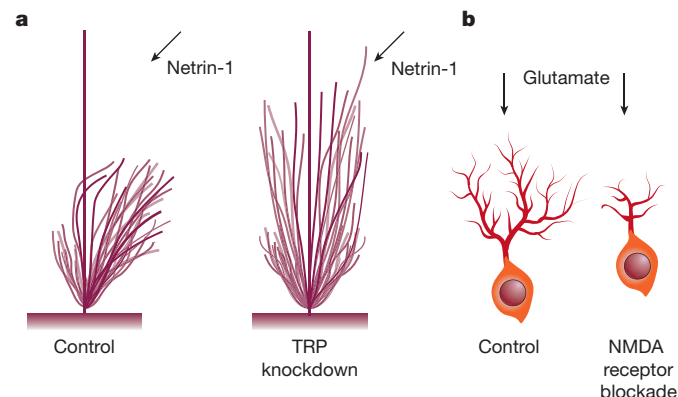


Figure 4 | Turning of embryonic axonal growth cones and elaboration of dendritic arbours depend on elevation of intracellular calcium. **a**, Growth cone trajectories of frog spinal neurons during exposure to a gradient of the guidance molecule netrin-1 (arrows) show turning in controls that is absent when synthesis of Ca-permeable TRP channels is knocked down. Adapted from ref. 38, with permission. **b**, Blocking Ca-permeable glutamate receptors (NMDAR) reduces the growth of dendrites in the frog brain. Adapted from ref. 49, with permission; copyright 1999 by the Society for Neuroscience.

Dendrite outgrowth. A fourth facet of neuronal differentiation is the formation of dendrites, which are generally shorter than axons, more extensively branched, and specialized for the receipt instead of transmission of synaptic signals. Their growth and stability depend on activation of neurotransmitter receptors and Ca elevation. Blockade of NMDA receptors on neurons in the *Xenopus* tadpole optic tectum, during the period of synaptogenesis in which axons are innervating the tectum and growth cones are releasing transmitter, suppresses development of their dendritic arbours by curbing addition of new branches and extension of pre-existing branches⁴⁹ (Fig. 4b). Cholinergic neurotransmission evokes local Ca-induced Ca release that stabilizes developing dendrites of retinal ganglion cells during early stages of synapse formation in the embryonic chick retina⁵⁰; blockade of glutamatergic transmission suppresses dendrite motility at later stages of development⁵¹. Dendrites of postnatal rat hippocampal neurons in slices generate Ca transients in their filopodia during synapse formation that are correlated with reduced filopodial motility⁵², similar to the effect of Ca transients on axonal filopodia⁴².

Activation of transmitter receptors also stimulates global Ca increases throughout dendritic arbours that signal to the nucleus and regulate dendrite outgrowth. Ca-induced dendritic growth in cultured embryonic rat cortical neurons is regulated by activation of a transcriptional program that involves CaM (Ca/calmodulin-dependent protein) kinase IV and CREB (cAMP responsive element binding protein)-mediated signalling to the nucleus⁵³. Targeted disruption of CREST, a Ca-responsive transactivator, compromises dendritic growth and branching in mouse hippocampal and cortical neurons⁵⁴. Thus activity-dependent transcriptional regulation via the activity of conventional transmitter-activated channels is a key aspect of dendrite formation. Activity of novel channels may also be involved, given the expression of TRP channels on dendrites in the mature nervous system.

Adult stem cells

The discovery of adult neurogenesis⁵⁵ led to investigation of the steps regulating proliferation, migration, and differentiation of newly born neurons as they develop in the adult brain. Remarkably, early stages of development of these neurons in the dentate gyrus of the hippocampus and in the olfactory bulb are also activity-dependent.

Neuronal precursors and differentiating neurons express voltage-gated Na and Ca and transmitter-activated channels similar to those in embryonic cells^{56,57}, which may give rise to spontaneous activity and have functions similar to those in the embryonic nervous system. Altering excitability affects proliferation. Activating NMDA receptors decreases proliferation and blocking NMDA receptors or lesioning excitatory afferents increases proliferation of granule cells in adult rat dentate gyrus⁵⁸. In contrast, serotonin stimulates hippocampal rodent neurogenesis^{59,60}. Seizure activity and voluntary wheel running increase the number of new neurons and odour deprivation reduces it^{61,62}, suggesting that physiological activity regulates adult neurogenesis. Perturbing excitability also perturbs migration. Seizures disrupt migration from the subgranular proliferative zone to the inner granule cell layer in the dentate gyrus⁶¹. Odour deprivation reduces expression of tenascin-R and disrupts radial migration of neuroblasts in the mouse olfactory bulb⁶³.

Depolarization by neurotransmitters is important for neuronal differentiation from adult stem cells. Excitation of cultured mouse hippocampal proliferating precursors via NMDA receptors and voltage-gated Ca channels represses glial fate genes and stimulates expression of the neural fate gene *NeuroD*⁶⁴. GABAergic synaptic activation of precursor cells *in vivo* is excitatory owing to their increased intracellular chloride concentration and stimulates Ca influx and expression of *NeuroD* that promotes neuronal differentiation⁶⁵. Progeny of adult rat hippocampal neural stem cells co-cultured with neurons and astrocytes from neonatal hippocampus become electrically active and form synaptic connections that appear

to contribute to their functional integration into pre-existing circuits⁶⁶. Indeed, integration of newly generated neurons into existing networks is activity-dependent. GABAergic depolarization of newborn mouse granule cells in the adult dentate gyrus is necessary for synapse formation and dendritic development *in vivo*⁶⁷. Thus many features of activity-dependent early development involving conventional ion channels are recapitulated in adult neurogenesis; roles for novel forms of activity remain to be investigated.

Conclusions and perspectives

Now that we recognize the involvement of electrical activity in many aspects of early neuronal development, we can ask whether electrical activity has a specific function or functions. Has evolution simply opportunistically seized on what was useful at the time—selecting signalling by transcription factors or signalling by ion channel activity for various purposes? Or are there particular benefits to the organism of signalling by ion channels that generate elevations of intracellular Ca?

On the one hand, signalling by electrical activity may refine signalling via gene expression, providing feedback loops to validate or fine-tune steps of development driven by genetic programs. Analogous to processes of development at later stages of differentiation, genes establish an initial landscape that subsequent activity reshapes. In this way electrical activity can act to ensure the efficiency and fidelity of brain assembly.

On the other hand, electrical activity seems likely to operate at a more fundamental level, integrated with gene expression. Genes encoding ion channels, receptors and ligands that activate them generate electrical activity leading to a wide range of elevations of intracellular Ca (ref. 68; Box 2). Different patterns of Ca elevation can regulate gene expression in a cell-autonomous manner. However, transient elevations of intracellular Ca also regulate cellular secretion and organization of the cytoskeleton that controls motility. Secretion of diffusible molecules and cellular motility that generates cell–cell contacts enable inductive interactions among cells that regulate neuronal development at all levels, from cell proliferation to post-translational protein modification. In this way brief electrical activity can act non-cell-autonomously to stimulate ligand–receptor binding with consequences that play out over longer, developmentally relevant timescales (Fig. 5).

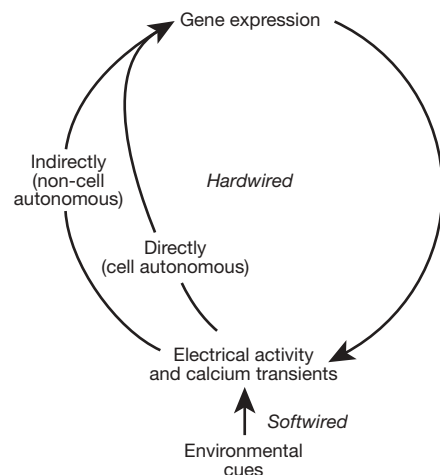


Figure 5 | The partnership of gene expression with electrical activity. Gene expression leads to synthesis of ion channels, receptors and ligands; secretion of ligands and cell–cell interactions drive electrical activity that generates transient elevations of Ca. Ca transients in turn regulate gene expression both directly in a cell-autonomous manner and indirectly, non-cell-autonomously, via inductive interactions. This hardwiring of early development is likely to be complemented by softwiring driven by stimulation of electrical activity by environmental cues.

Box 2 | Spontaneous activity following synapse formation

Spontaneous activity is also observed after synapse formation. Ca transients generated at these later stages of development may influence aspects of differentiation (for example, neurotransmitter specification) that are regulated by early spontaneous activity and affect development of neighbouring neurons that are at earlier stages of development. Rhythmic spontaneous electrical activity and Ca transients are observed early in embryonic mouse and chick spinal cord development^{82,83}. Voltage-sensitive dyes allow optical identification of transiently expressed Ca-dependent action potentials and spontaneous large-scale correlated wave activity in the embryonic chick and rat brainstem^{84,85}. Synchronized Ca transients stimulated by serotonin are observed in mouse cranial motor neurons at embryonic stages of development⁸⁶. Synchronized Ca oscillations activated by AMPA and NMDA receptors are propagated from posterior to anterior cortex in the newborn rat⁸⁷. Similar transients are generated by voltage-gated Na and Ca channels in mouse cortex and the degree of synchrony peaks at the time of birth⁸⁸. Postnatally, spontaneous giant depolarizing potentials are observed in synchronized populations of rat hippocampal neurons⁸⁹ that generate Ca transients⁹⁰. Developing retinal ganglion cells propagate spontaneous bursts of action potentials and ensuing Ca elevations in waves across the retina^{91,92}. Spontaneous synchronous and sequential Ca transients driven by intrinsic currents and circuit mechanisms are observed in mouse visual cortex in networks of pyramidal cells⁹³, with precise repetitions of sequences of activation in mouse and cat cortex⁹⁴. Recurrent Ca transients involving thousands of cells occur in cortical neurons of non-aesthetized newborn mice during resting periods⁹⁵. It is of great interest to understand further the functions of these forms of spontaneous activity.

This process is probably important, because the complexity of the brain renders the number of genes in the genome insufficient to program the organization of the nervous system directly on a single-gene-to-single-component basis. Local communication is required to specify developmental choices in a temporally and spatially appropriate manner. Cellular and subcellular Ca signalling controls DNA synthesis^{6–9,58–62,64,65}, regulates messenger RNA and protein levels^{47,48,54,69}, and stimulates phosphorylation and dephosphorylation^{53,70,71}. Thus electrical activity and gene expression appear to be full partners in neuronal development. We can look ahead to identification of the inducing molecules released in an activity-dependent manner and the cell–cell contacts driven by activity that drive inductive interactions in early neuronal development.

Activity-dependence of early neural development does not necessarily challenge the idea that development is hardwired. Development operates on a fixed agenda if activity simply implements the outcome of genetic programs or even if it is an equal partner, turning on and being turned on by genetic programs. Hardwiring will be broken, however, when environmental cues alter activity that affects development (Fig. 5). Because synapses have not yet formed at early stages, these cues must reach developing neurons by other means, such as paracrine action of transmitters, circulation of maternal hormones, or (in lower animals) fluctuations in temperature. These stimuli can be predicted to alter early activity, but the extent to which this regulates and softwires early development remains to be determined.

1. Zito, K. & Svoboda, K. Activity-dependent synaptogenesis in the adult mammalian cortex. *Neuron* **35**, 1015–1017 (2002).
2. Goda, Y. & Davis, G. W. Mechanisms of synapse assembly and disassembly. *Neuron* **40**, 243–264 (2003).
3. Mennerick, S. & Zorumski, C. F. Neural activity and survival in the developing nervous system. *Mol. Neurobiol.* **22**, 41–54 (2000).
4. Vanhoutte, P. & Bading, H. Opposing roles of synaptic and extrasynaptic NMDA receptors in neuronal calcium signalling and BDNF gene regulation. *Curr. Opin. Neurobiol.* **13**, 366–371 (2003).
5. Webb, S. E., Moreau, M., Leclerc, C. & Miller, A. L. Calcium transients and neural induction in vertebrates. *Cell Calcium* **37**, 375–385 (2005).

6. LoTurco, J. J., Owens, D. F., Heath, M. J., Davis, M. B. & Kriegstein, A. R. GABA and glutamate depolarize cortical progenitor cells and inhibit DNA synthesis. *Neuron* **15**, 1287–1298 (1995).
7. Liu, X., Wang, Q., Haydar, T. F. & Bordey, A. Nonsynaptic GABA signaling in postnatal subventricular zone controls proliferation of GFAP-expressing progenitors. *Nature Neurosci.* **8**, 1179–1187 (2005).
8. Fiszman, M. L., Borodinsky, L. N. & Neale, J. H. GABA induces proliferation of immature cerebellar granule cells grown *in vitro*. *Dev. Brain Res.* **115**, 1–8 (1999).
9. Weissman, T. A., Riquelme, P. A., Ivic, L., Flint, A. C. & Kriegstein, A. R. Calcium waves propagate through radial glial cells and modulate proliferation in the developing neocortex. *Neuron* **43**, 647–661 (2004).
10. Erhardt, N. M. & Sherwood, N. M. PACAP maintains cell cycling and inhibits apoptosis in chick neuroblasts. *Mol. Cell. Endocrinol.* **221**, 121–134 (2004).
11. Gibbs, S. M. Regulation of neuronal proliferation and differentiation by nitric oxide. *Mol. Neurobiol.* **27**, 107–120 (2003).
12. Komuro, H. & Rakic, P. Intracellular Ca²⁺ fluctuations modulate the rate of neuronal migration. *Neuron* **17**, 275–285 (1996).
13. Demarque, M. *et al.* Paracrine intercellular communication by a Ca²⁺- and SNARE-independent release of GABA and glutamate prior to synapse formation. *Neuron* **36**, 1051–1061 (2002).
14. Manent, J. B. *et al.* A noncanonical release of GABA and glutamate modulates neuronal migration. *J. Neurosci.* **25**, 4755–4765 (2005).
15. Yacubova, E. & Komuro, H. Stage-specific control of neuronal migration by somatostatin. *Nature* **415**, 77–81 (2002).
16. Bolteus, A. J. & Bordey, A. GABA release and uptake regulate neuronal precursor migration in the postnatal subventricular zone. *J. Neurosci.* **24**, 7623–7631 (2004).
17. Chen, Y. *et al.* Reelin modulates NMDA receptor activity in cortical neurons. *J. Neurosci.* **25**, 8209–8216 (2005).
18. Haase, A. & Bicker, G. Nitric oxide and cyclic nucleotides are regulators of neuronal migration in an insect embryo. *Development* **130**, 3977–3987 (2003).
19. Dallman, J. E., Davis, A. K. & Moody, W. J. Spontaneous activity regulates Ca²⁺-dependent K⁺ current expression in developing ascidian muscle. *J. Physiol. (Lond.)* **511**, 683–693 (1998).
20. Gu, X. & Spitzer, N. C. Distinct aspects of neuronal differentiation encoded by frequency of spontaneous Ca²⁺ transients. *Nature* **375**, 784–787 (1995).
21. Desarmenien, M. G. & Spitzer, N. C. Role of Ca and protein kinase C in development of the delayed rectifier potassium current in *Xenopus* spinal neurons. *Neuron* **7**, 797–805 (1991).
22. Borodinsky, L. N. *et al.* Activity-dependent homeostatic specification of transmitter expression in embryonic neurons. *Nature* **429**, 523–530 (2004).
23. Ahmed, Z., Walker, P. S. & Fellows, R. E. Properties of neurons from dissociated fetal rat brain in serum-free culture. *J. Neurosci.* **3**, 2448–2462 (1983).
24. Nerbonne, J. M. & Gurney, A. M. Development of excitable membrane properties in mammalian sympathetic neurons. *J. Neurosci.* **9**, 3272–3286 (1989).
25. McCobb, D. P., Best, P. M. & Beam, K. G. The differentiation of excitability in embryonic chick limb motoneurons. *J. Neurosci.* **10**, 2974–2984 (1990).
26. Brosenitsch, T. A. & Katz, D. M. Physiological patterns of electrical stimulation can induce neuronal gene expression by activating N-type calcium channels. *J. Neurosci.* **21**, 2571–2579 (2001).
27. Brosenitsch, T. A. & Katz, D. M. Expression of Phox2 transcription factors and induction of the dopaminergic phenotype in primary sensory neurons. *Mol. Cell. Neurosci.* **20**, 447–457 (2002).
28. Nishimaru, H., Restrepo, C. E., Ryge, J., Yanagawa, Y. & Kiehn, O. Mammalian motor neurons corelease glutamate and acetylcholine at central synapses. *Proc. Natl Acad. Sci. USA* **102**, 5245–5249 (2005).
29. Mentis, G. Z. *et al.* Noncholinergic excitatory actions of motoneurons in the neonatal mammalian spinal cord. *Proc. Natl Acad. Sci. USA* **102**, 7344–7349 (2005).
30. Brunelli, G. *et al.* Glutamatergic reinnervation through peripheral nerve graft dictates assembly of glutamatergic synapses at rat skeletal muscle. *Proc. Natl Acad. Sci. USA* **102**, 8752–8757 (2005).
31. Gomez, T. M., Snow, D. M. & Letourneau, P. C. Characterization of spontaneous Ca transients in nerve growth cones and their effect on growth cone migration. *Neuron* **14**, 1233–1246 (1995).
32. Gomez, T. M. & Spitzer, N. C. *In vivo* regulation of axon extension and pathfinding by growth-cone Ca transients. *Nature* **397**, 350–355 (1999).
33. Tang, F., Dent, E. W. & Kalil, K. Spontaneous calcium transients in developing cortical neurons regulate axon outgrowth. *J. Neurosci.* **23**, 927–936 (2003).
34. Greka, A., Navarro, B., Oancea, E., Duggan, A. & Clapham, D. E. TRPC5 is a regulator of hippocampal neurite length and growth cone morphology. *Nature Neurosci.* **6**, 837–845 (2003).
35. Zheng, J. Q., Wan, J. J. & Poo, M. M. Essential role of filopodia in chemotropic turning of nerve growth cone induced by a glutamate gradient. *J. Neurosci.* **16**, 1140–1149 (1996).
36. Hong, K., Nishiyama, M., Henley, J., Tessier-Lavigne, M. & Poo, M. Calcium signalling in the guidance of nerve growth by netrin-1. *Nature* **403**, 93–98 (2000).
37. Nishiyama, M. *et al.* Cyclic AMP/GMP-dependent modulation of Ca²⁺ channels sets the polarity of nerve growth-cone turning. *Nature* **423**, 990–995 (2003).
38. Wang, G. X. & Poo, M. Requirement of TRPC channels in netrin-1-induced chemotropic turning of nerve growth cones. *Nature* **434**, 898–904 (2005).
39. Shim, S. *et al.* XTRPC1-dependent chemotropic guidance of neuronal growth cones. *Nature Neurosci.* **8**, 730–735 (2005).

40. Li, Y. *et al.* Essential role of TRPC channels in the guidance of nerve growth cones by brain-derived neurotrophic factor. *Nature* **434**, 894–898 (2005).
41. Ming, G., Henley, J., Tessier-Lavigne, M., Song, H. & Poo, M. Electrical activity modulates growth cone guidance by diffusible factors. *Neuron* **29**, 441–452 (2001).
42. Gomez, T. M., Robles, E., Poo, M. & Spitzer, N. C. Filopodial calcium transients promote substrate-dependent growth cone turning. *Science* **291**, 1983–1987 (2001).
43. Ruthazer, E. S., Akerman, C. J. & Cline, H. T. Control of axon branch dynamics by correlated activity *in vivo*. *Science* **301**, 66–70 (2003).
44. Hua, J. Y., Smear, M. C., Baier, H. & Smith, S. J. Regulation of axon growth *in vivo* by activity-based competition. *Nature* **434**, 1022–1026 (2005).
45. O'Leary, D. D. *et al.* A target-derived chemoattractant controls the development of the corticopontine projection by a novel mechanism of axon targeting. *Dev., Suppl.* **2**, 123–130 (1991).
46. Tang, F. & Kalil, K. Netrin-1 induces axon branching in developing cortical neurons by frequency-dependent calcium signaling pathways. *J. Neurosci.* **25**, 6702–6715 (2005).
47. Itoh, K., Stevens, B., Schachner, M. & Fields, R. D. Regulated expression of the neural cell adhesion molecule L1 by specific patterns of neural impulses. *Science* **270**, 1369–1372 (1995).
48. Hanson, M. G. & Landmesser, L. T. Normal patterns of spontaneous activity are required for correct motor axon guidance and the expression of specific guidance molecules. *Neuron* **43**, 687–701 (2004).
49. Rajan, I. & Cline, H. T. Glutamate receptor activity is required for normal development of tectal cell dendrites *in vivo*. *J. Neurosci.* **18**, 7836–7846 (1998).
50. Lohmann, C., Myhr, K. L. & Wong, R. O. Transmitter-evoked local calcium release stabilizes developing dendrites. *Nature* **418**, 177–181 (2002).
51. Wong, W. T., Faulkner-Jones, B. E., Sanes, J. R. & Wong, R. O. Rapid dendritic remodeling in the developing retina: dependence on neurotransmission and reciprocal regulation by Rac and Rho. *J. Neurosci.* **20**, 5024–5036 (2000).
52. Lohmann, C., Finski, A. & Bonhoeffer, T. Local Ca transients regulate the spontaneous motility of dendritic filopodia. *Nature Neurosci.* **8**, 305–312 (2005).
53. Redmond, L., Kashani, A. H. & Ghosh, A. Calcium regulation of dendritic growth via CaM kinase IV and CREB-mediated transcription. *Neuron* **34**, 999–1010 (2002).
54. Aizawa, H. *et al.* Dendrite development regulated by CREST, a calcium-regulated transcriptional activator. *Science* **303**, 197–202 (2004).
55. Altman, J. & Das, G. D. Autoradiographic and histological evidence of postnatal hippocampal neurogenesis in rats. *J. Comp. Neurol.* **124**, 319–335 (1965).
56. Schmidt-Hieber, C., Jonas, P. & Bischofberger, J. Enhanced synaptic plasticity in newly generated granule cells of the adult hippocampus. *Nature* **429**, 184–187 (2004).
57. Esposito, M. S. *et al.* Neuronal differentiation in the adult hippocampus recapitulates embryonic development. *J. Neurosci.* **25**, 10074–10086 (2005).
58. Cameron, H. A., McEwen, B. S. & Gould, E. Regulation of adult neurogenesis by excitatory input and NMDA receptor activation in the dentate gyrus. *J. Neurosci.* **15**, 4687–4692 (1995).
59. Malberg, J. E., Eisch, A. J., Nestler, E. J. & Duman, R. S. Chronic antidepressant treatment increases neurogenesis in adult rat hippocampus. *J. Neurosci.* **20**, 9104–9110 (2000).
60. Santarelli, L. *et al.* Requirement of hippocampal neurogenesis for the behavioral effects of antidepressants. *Science* **301**, 805–809 (2003).
61. Parent, J. M. *et al.* Dentate granule cell neurogenesis is increased by seizures and contributes to aberrant network reorganization in the adult rat hippocampus. *J. Neurosci.* **17**, 3727–3738 (1997).
62. van Praag, H., Christie, B. R., Sejnowski, T. J. & Gage, F. H. Running enhances neurogenesis, learning, and long-term potentiation in mice. *Proc. Natl Acad. Sci. USA* **96**, 13427–13431 (1999).
63. Saghatelian, A., de Chevigny, A., Schachner, M. & Lledo, P. M. Tenascin-R mediates activity-dependent recruitment of neuroblasts in the adult mouse forebrain. *Nature Neurosci.* **7**, 347–356 (2004).
64. Deisseroth, K. *et al.* Excitation-neurogenesis coupling in adult neural stem/progenitor cells. *Neuron* **42**, 535–552 (2004).
65. Tozuka, Y., Fukuda, S., Namba, T., Seki, T. & Hisatsune, T. GABAergic excitation promotes neuronal differentiation in adult hippocampal progenitor cells. *Neuron* **47**, 803–815 (2005).
66. Song, H. J., Stevens, C. F. & Gage, F. H. Neural stem cells from adult hippocampus develop essential properties of functional CNS neurons. *Nature Neurosci.* **5**, 438–445 (2002).
67. Ge, S. *et al.* GABA regulates synaptic integration of newly generated neurons in the adult brain. *Nature* **439**, 589–593 (2006).
68. Berridge, M. J., Bootman, M. D. & Roderick, H. L. Calcium signalling: dynamics, homeostasis and remodeling. *Nature Rev. Mol. Cell Biol.* **4**, 517–529 (2003).
69. Watt, S. D., Gu, X., Smith, R. D. & Spitzer, N. C. Specific frequencies of spontaneous Ca^{2+} transients upregulate GAD 67 transcripts in embryonic spinal neurons. *Mol. Cell. Neurosci.* **16**, 376–387 (2000).
70. Wu, G. Y. & Cline, H. T. Stabilization of dendritic arbor structure *in vivo* by CaMKII. *Science* **279**, 222–226 (1998).
71. Lautermilch, N. J. & Spitzer, N. C. Regulation of calcineurin by growth cone calcium waves controls neurite extension. *J. Neurosci.* **20**, 315–325 (2000).
72. Blanton, M. G., Lo Turco, J. J. & Kriegstein, A. R. Endogenous neurotransmitter activates N-methyl-D-aspartate receptors on differentiating neurons in embryonic cortex. *Proc. Natl Acad. Sci. USA* **87**, 8027–8030 (1990).
73. Flint, A. C., Dammerman, R. S. & Kriegstein, A. R. Endogenous activation of metabotropic glutamate receptors in neocortical development causes neuronal Ca oscillations. *Proc. Natl Acad. Sci. USA* **96**, 12144–12149 (1999).
74. Wang, D. D., Krueger, D. D. & Bordey, A. GABA depolarizes neuronal progenitors of the postnatal subventricular zone via GABA_A receptor activation. *J. Physiol. (Lond.)* **550**, 785–800 (2003).
75. Carey, M. B. & Matsumoto, S. G. Spontaneous calcium transients are required for neuronal differentiation of murine neural crest. *Dev. Biol.* **215**, 298–313 (1999).
76. Gu, X., Olson, E. C. & Spitzer, N. C. Spontaneous neuronal calcium spikes and waves during early differentiation. *J. Neurosci.* **14**, 6325–6335 (1994).
77. Conklin, M. W., Lin, M. S. & Spitzer, N. C. Local calcium transients contribute to disappearance of pFAK, focal complex removal and deadhesion of neuronal growth cones and fibroblasts. *Dev. Biol.* **287**, 201–212 (2005).
78. Ashworth, R. & Bolsover, S. R. Spontaneous activity-independent intracellular calcium signals in the developing spinal cord of the zebrafish embryo. *Dev. Brain Res.* **139**, 131–137 (2002).
79. Owens, D. F. & Kriegstein, A. R. Patterns of intracellular calcium fluctuation in precursor cells of the neocortical ventricular zone. *J. Neurosci.* **18**, 5374–5388 (1998).
80. Mercer, A. R. & Hildebrand, J. G. Developmental changes in the electrophysiological properties and response characteristics of *Manduca* antennal-lobe neurons. *J. Neurophysiol.* **87**, 2650–2663 (2002).
81. Jiang, S. A., Campusano, J. M., Su, H. & O'Dowd, D. K. *Drosophila* mushroom body Kenyon cells generate spontaneous calcium transients mediated by PLTX-sensitive calcium channels. *J. Neurophysiol.* **94**, 491–500 (2005).
82. Hanson, M. G. & Landmesser, L. T. Characterization of the circuits that generate spontaneous episodes of activity in the early embryonic mouse spinal cord. *J. Neurosci.* **23**, 587–600 (2003).
83. O'Donovan, M. J., Bonnot, A., Wenner, P. & Mentis, G. Z. Calcium imaging of network function in the developing spinal cord. *Cell Calcium* **37**, 443–450 (2005).
84. Momose-Sato, Y., Sato, K. & Kamino, K. Optical identification of Ca-dependent action potentials transiently expressed in the embryonic rat brainstem. *Neuroscience* **90**, 1293–1310 (1999).
85. Momose-Sato, Y., Mochida, H., Sasaki, S. & Sato, K. Depolarization waves in the embryonic CNS triggered by multiple sensory inputs and spontaneous activity: optical imaging with a voltage-sensitive dye. *Neuroscience* **116**, 407–423 (2003).
86. Hunt, P. N., McCabe, A. K. & Bosma, M. M. Midline serotonergic neurones contribute to widespread synchronized activity in embryonic mouse hindbrain. *J. Physiol. (Lond.)* **566**, 807–819 (2005).
87. Garaschuk, O., Linn, J., Eilers, J. & Konnerth, A. Large-scale oscillatory calcium waves in the immature cortex. *Nature Neurosci.* **3**, 452–459 (2000).
88. Corlew, R., Bosma, M. M. & Moody, W. J. Spontaneous, synchronous electrical activity in neonatal mouse cortical neurones. *J. Physiol. (Lond.)* **560**, 377–390 (2004).
89. Ben-Ari, Y., Cherubini, E., Corradetti, R. & Gaiarsa, J. L. Giant synaptic potentials in immature rat CA3 hippocampal neurones. *J. Physiol. (Lond.)* **416**, 303–325 (1989).
90. Kasyanov, A. M., Safiulina, V. F., Voronin, L. L. & Cherubini, E. GABA-mediated giant depolarizing potentials as coincidence detectors for enhancing synaptic efficacy in the developing hippocampus. *Proc. Natl Acad. Sci. USA* **101**, 3967–3972 (2004).
91. Stellwagen, D. & Shatz, C. J. An instructive role for retinal waves in the development of retinogeniculate connectivity. *Neuron* **33**, 357–367 (2002).
92. Torborg, C. L., Hansen, K. A. & Feller, M. B. High frequency, synchronized bursting drives eye-specific segregation of retinogeniculate projections. *Nature Neurosci.* **8**, 72–78 (2005).
93. Cossart, R., Aronov, D. & Yuste, R. Attractor dynamics of network UP states in the neocortex. *Nature* **423**, 283–288 (2003).
94. Ikegaya, Y. *et al.* Synfire chains and cortical songs: temporal modules of cortical activity. *Science* **304**, 559–564 (2004).
95. Adelsberger, H., Garaschuk, O. & Konnerth, A. Cortical calcium waves in resting newborn mice. *Nature Neurosci.* **8**, 988–990 (2005).
96. Borodinsky, L. N. & Spitzer, N. C. Activity-dependent neurotransmitter–receptor matching at the neuromuscular junction. *Proc. Natl Acad. Sci. USA*, in the press.

Acknowledgements I thank D. Berg, L. Borodinsky, M. Feller, A. Ghosh and K. Marek for comments on the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The author declares no competing financial interests. Correspondence should be addressed to the author (nsptizer@ucsd.edu).

Insights into the dynamics of mantle plumes from uranium-series geochemistry

Bernard Bourdon^{1†}, Neil M. Ribe², Andreas Stracke^{3†}, Alberto E. Saal⁴ & Simon P. Turner⁵

The long-standing paradigm that hotspot volcanoes such as Hawaii or Iceland represent the surface expression of mantle plumes—hot, buoyant upwelling regions beneath the Earth's lithosphere—has recently been the focus of controversy. Whether mantle plumes exist or not is pivotal for our understanding of the thermal, dynamic and compositional evolution of the Earth's mantle. Here we show that uranium-series disequilibria measured in hotspot lavas indicate that hotspots are indeed associated with hot and buoyant upwellings and that weaker (low buoyancy flux) hotspots such as Iceland and the Azores are characterized by lower excess temperatures than stronger hotspots such as Hawaii. This direct link between buoyancy flux and mantle temperature is evidence for the existence of mantle plumes.

The existence of mantle plumes has recently been questioned on the basis of geophysical, petrological and geochemical arguments^{1,2}. Following McKenzie³, we use the term 'plume' in a strictly fluid-dynamical sense, that is "a buoyant upwelling or downwelling whose buoyancy results from the material in the plume being hotter or colder than the surrounding mantle with no implications whatsoever about the depth to which the circulation extends, or about whether or not relative motion between different plumes occurs, or whether the thermal buoyancy is associated with compositional or isotopic variations". According to this definition, evidence for the existence of mantle plumes should come mainly from fluid-dynamical arguments. The persistent difficulty of demonstrating that hotspots are the surface expressions of mantle plumes is due to a number of factors: the poor resolution of geophysical methods, the inability of geophysical and petrological data to show unambiguously that the mantle beneath hot spots is anomalously hot, and the absence of direct geophysical constraints on the velocity of the upwelling mantle.

Recently, however, it has been shown that U-series disequilibria in young hotspot lavas^{4–8} provide a relative measure of mantle upwelling velocity beneath ocean islands. Thus, geochemical data provide important complementary information that is not available from geophysical and petrological data alone. Admittedly, constraining mantle upwelling velocities with U-series data requires that the influence of inter-hotspot differences in the major-element composition of the mantle source during melt production can either be corrected for^{9–11} or is relatively minor. While the role of source heterogeneity on U-series has sometimes been clearly identified (for example, the Sao Miguel island in the Azores region⁶), Stracke *et al.*¹² argue that the role of source heterogeneity on trace-element partitioning during melting beneath ocean islands is subordinate relative to other effects such as variation in mantle upwelling velocity. However, these authors¹² also show that the large variability in mineral/melt partition coefficients for U-series nuclides precludes any definite conclusion on the effect of source heterogeneity⁶ on partitioning behaviour. Accordingly, in the models presented below we assume that the effects of source heterogeneity can be neglected relative to the effects of other key parameters. Our results demonstrate that U-series

data in ocean island lavas provide constraints on mantle upwelling velocity, mantle temperatures (which control the degree of melting) and the horizontal length scale of mantle upwelling, thereby establishing U-series data as a unique link between geochemical and geophysical constraints on mantle dynamics.

U-series systematics in ocean island basalts

The database we have assembled includes new U–Th–Pa mass spectrometry data from the Azores^{6,13,14}, Pitcairn (B.B. *et al.*, manuscript in preparation), the Galapagos islands (A.E.S. *et al.*, manuscript in preparation), and Iceland^{11,12} in addition to published data from Hawaii^{8,15}, the Canary islands¹⁶, the Afar region¹⁷ and Iceland⁷. Other data are from a compilation by Chabaux and Allègre¹⁸. All the samples we consider are lavas from either historical or dated eruptions, so that the effect of radioactive decay since eruption can be either neglected or corrected for. The measurements upon which our analysis is based are the 'activity' ratios (²³⁰Th/²³⁸U) and (²³¹Pa/²³⁵U), where the 'activity' of each nuclide (²³⁰Th, ²³⁸U, ²³¹Pa, or ²³⁵U) is the product λN of its radioactive decay constant λ and its population (number of atoms) N in the sample. As a consequence of the exponential law for radioactive decay, the activity ratio for any two nuclides in a given series is equal to unity if the system is in 'secular equilibrium' (that is, there is a constant population of each nuclide), as is the case in the mantle source before melting. Activity ratios (²³⁰Th/²³⁸U) and (²³¹Pa/²³⁵U) in basalts that differ from unity therefore provide a direct measure of melting-induced fractionation between U–Th and U–Pa, respectively.

On a global scale, there are some remarkable correlations between the U-series data in hotspot lavas and two important geophysical parameters: the buoyancy flux B of the hotspot (essentially the product of the speed, the cross-sectional area, and the density anomaly of the upwelling) and the distance r from the sample to the centre of the hotspot. First, there is a negative correlation between excess ²³⁰Th in the lavas and B (Fig. 1a), first noted by Chabaux and Allègre¹⁸, who proposed that it reflects variations in the degree of melting (see discussion below). Our new data reveal that there is a similar inverse correlation between ²³¹Pa excess and B (Fig. 1b).

¹Laboratoire de Géochimie et Cosmochimie, Institut de Physique du Globe de Paris-CNRS, 4 Place Jussieu, 75252 Paris cedex 05, France. ²Laboratoire de Dynamique des Systèmes Géologiques, Institut de Physique du Globe de Paris-CNRS, 4 Place Jussieu, 75252 Paris cedex 05, France. ³Max-Planck-Institut für Chemie, Abteilung Geochemie, Postfach 3060, 55020 Mainz Germany. ⁴Department of Geological Sciences, Brown University, Providence, Rhode Island 02912, USA. ⁵GEMOC, Department of Earth and Planetary Sciences, Macquarie University, Sydney, New South Wales 2109, Australia. [†]Present addresses: Institute of Isotope Geochemistry and Mineral Resources, Department of Earth Sciences, ETH Zürich, CH-8092 Zürich, Switzerland.

Second, parent–daughter activity ratios (for example, $^{231}\text{Pa}/^{235}\text{U}$, $^{230}\text{Th}/^{238}\text{U}$) show a positive trend with r for several hotspots, including Hawaii⁸, Iceland⁷ and $^{231}\text{Pa}/^{235}\text{U}$ in the Azores⁶ (Fig. 2) and the Galapagos (A.E.S. *et al.*, manuscript in preparation). In these diagrams, the centre of the hotspot is estimated from geophysical data for Iceland¹⁹ and Hawaii²⁰ and from He isotope data²¹ for the Azores (Fig. 2a). (By contrast, there is no clear trend for the Canaries, where plume–lithosphere interaction may affect the U-series systematics¹⁶). The correlations shown in Figs 1 and 2 are intriguing because they directly relate the geochemical characteristics of erupted lavas to geophysical parameters of the underlying mantle (for example, mantle temperature, upwelling velocity, and possibly the degree of melting¹⁸). We now propose a model to quantify the influence of these parameters on the U-series systematics of erupted lavas.

Inferences of temperature and upwelling velocity beneath hotspots

To model theoretically the relationship between U-series disequilibrium and buoyancy flux (Fig. 1), we make two reasonable simplifying assumptions. First, we suppose that the melt productivity (percent-

age melt per kbar) is constant, which implies that the mantle upwelling velocity is linearly related to the melting rate (in kg of melt per cubic metre per year). Second, we assume that there exists a simple quantitative relationship between B and the mantle upwelling velocity W (for example, $B \propto W^2$; ref. 4). In this case, melting models^{5,18,22} that include the melting rate and assume that melt is extracted from the matrix once a critical porosity is reached predict that activity ratios (such as $^{231}\text{Pa}/^{235}\text{U}$ or $^{230}\text{Th}/^{238}\text{U}$) in the melt should be inversely related to the upwelling velocity (Fig. 3). On this basis, we used the dynamic melting equation given in ref. 22 to calculate the activity ratios ($^{231}\text{Pa}/^{235}\text{U}$) or ($^{230}\text{Th}/^{238}\text{U}$) in the melt as a function of buoyancy flux, assuming that the melting rate, excess temperature, viscosity and total extent of melting are all constant. This model predicts that ($^{231}\text{Pa}/^{235}\text{U}$) in the melt decreases as B increases, reaching a value of unity for very large B . This first-order approach supports the idea that variations in the magnitude of U-series disequilibrium are mainly controlled by variations in mantle upwelling rates. This differs from the conclusion of ref. 18 that the controlling factor is the degree of melting F_{max} . However, variations in F_{max} , while they may cause some dispersion in the observed relationships, cannot explain the overall trend shown on Fig. 1, because the Hawaii, Iceland and Afar samples include both high-degree and low-degree melts (tholeiites and alkali basalts, respectively) while the Canaries, Azores, Samoa and Pitcairn samples are exclusively low-degree melts (alkali basalts).

Although the simple melting model used above lends plausibility to the hypothesis that U-series systematics are controlled by the upwelling velocity, it neglects important thermal aspects of the problem such as differences of mantle temperature beneath different hotspots, the temperature-dependence of mantle viscosity, and the fact that weaker plumes (those with lower B) cool more than stronger plumes as they ascend²³. Accordingly, we now test the robustness of our preliminary conclusion using a more sophisticated plume model²⁴ that properly incorporates thermal effects. According to this model, the excess temperature ΔT_{top} , at the top of a plume (that is, just beneath the lithosphere) can be calculated as a function of the excess temperature ΔT_{bottom} at the plume's depth of origin by:

$$\Delta T_{\text{top}} = \Delta T_{\text{bottom}} \exp \left[- \frac{4\pi k \alpha}{\beta C_p B} z \right] \quad (1)$$

where B is the buoyancy flux, β is a parameter describing the dependence of viscosity on temperature, and the remaining parameters are defined in Supplementary Table 1. Although equation (1) neglects the dependence of thermal conductivity (k) and thermal expansivity (α) on pressure, its predictions are similar to those of more realistic models²³, and it can therefore be used to calculate the relationships among B , ΔT_{top} and ΔT_{bottom} (Supplementary Fig. 1). For a given buoyancy flux B , ΔT_{top} depends strongly on ΔT_{bottom} , except for low buoyancy fluxes where ΔT_{top} is relatively low and less dependent on the initial ΔT_{bottom} . According to ref. 25, mantle plumes may be of different types with different values of ΔT_{bottom} , depending on the depth where the plume originates and on the volume of the hot boundary layer that it samples. In applying equation (1), therefore, we treated ΔT_{bottom} and B as variable input parameters and then calculated the U-series signature in the erupted melt using a dynamic melting model (see Supplementary information). The model accounts for the fact that upwelling material that is hotter begins to melt at a greater depth, and also (by extension) for the fact that the degree of melting is larger for greater initial depths of melting.

The results of the calculations (Fig. 4) show that the observations are well explained by moderate (50–200 °C) values of the top excess temperature ΔT_{top} . According to Supplementary Fig. 1, these values of ΔT_{top} correspond to values of ΔT_{bottom} that are much smaller than the temperature jump across the boundary layer at the core–mantle boundary, which may be as much as 1,300 °C (ref. 26). This would indicate that only the upper part of the thermal boundary layer is

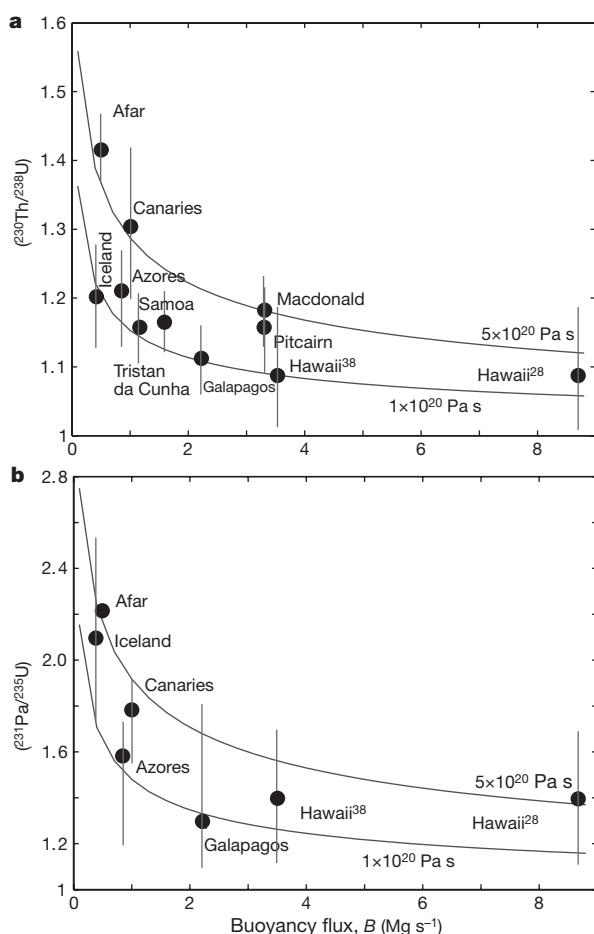


Figure 1 | U-series activity ratios versus buoyancy fluxes for recent hotspot lavas. $^{230}\text{Th}/^{238}\text{U}$ (a) and $^{231}\text{Pa}/^{235}\text{U}$ (b) activity ratios versus buoyancy fluxes for recent hotspot lavas. The mean for each hotspot is shown by a black circle and the range by a vertical line. Both diagrams show a negative trend, despite relatively large uncertainties in the estimation of buoyancy fluxes. This trend can be explained by variation in upwelling velocities. Other possibilities are discussed in detail in the text. Buoyancy fluxes are taken from ref. 38 when available, except for Hawaii where the estimates of both ref. 38 and ref. 28 are given. Curves in a and b are labelled with viscosity at the axis of the plume. The model curves were calculated with a dynamic melting model with constant melt productivity, constant viscosity and excess temperature (see Supplementary information for description of the model).

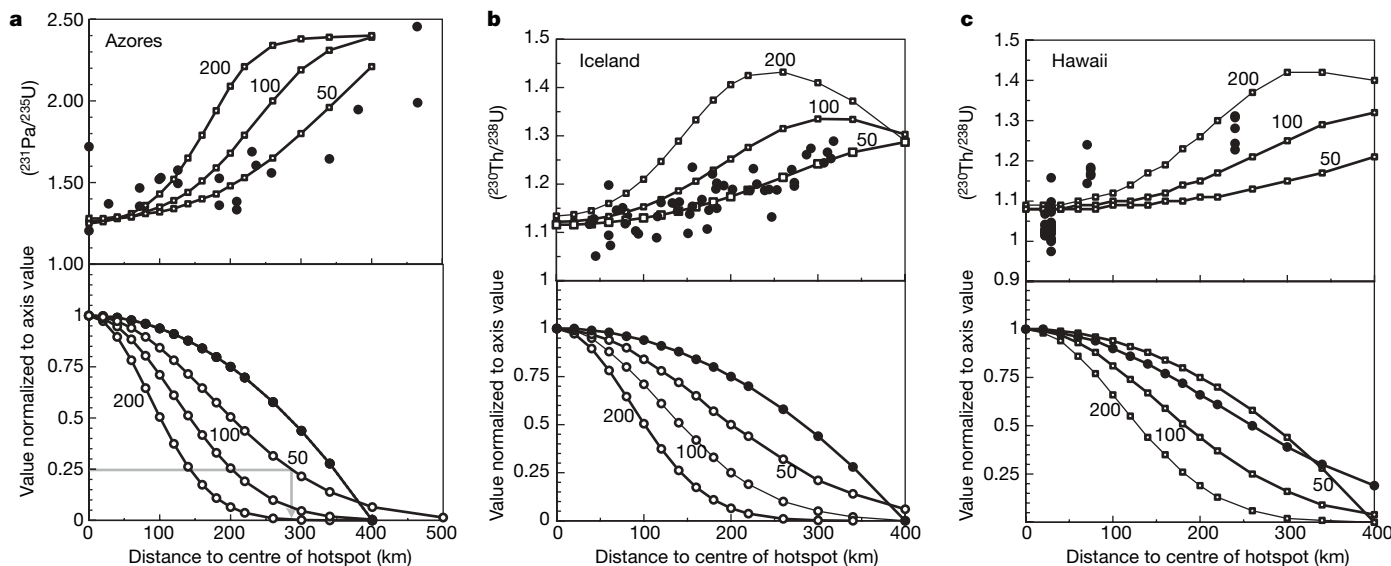


Figure 2 | U-series activity ratios as a function of distance from the centre of hotspots. **a**, $(^{231}\text{Pa}/^{235}\text{U})$ activity ratios plotted as a function of the distance from the centre of the Azores hotspot (data from ref. 6). **b**, $(^{230}\text{Th}/^{238}\text{U})$ activity ratios plotted as a function of the distance from the centre of the Iceland hotspot (data from ref. 7). **c**, $(^{230}\text{Th}/^{238}\text{U})$ activity ratios plotted as a function of the distance from the centre of the Hawaii hotspot (data as in ref. 5). The curves are calculated using a dynamic melting model and an analytical model for an axisymmetric plume²⁴. The bottom panels show the profiles of excess temperature (black symbols) and upwelling velocity (white symbols) predicted by the model, normalized to their

maximum values at the plume axis. The excess temperature at the top of the plume relative to the ambient mantle (numbers on curves) was adjusted to give the best fit to the observations. The best fit to the Hawaii data was estimated to be 200 °C based on the combined ^{230}Th and ^{231}Pa data set (not shown). The degree of melting used in the model is a function of the initial melting temperature. The melts are assumed to move vertically with no lateral mixing. The radius of the plume (r_{25}) is defined by the distance from the centre of the plume where $W = 0.25W_0$, where W_0 is the axial upwelling velocity.

being sampled by those plumes that originate from the core–mantle boundary²⁷.

The calculations also show that the dependence of $(^{231}\text{Pa}/^{235}\text{U})$ and $(^{230}\text{Th}/^{238}\text{U})$ on ΔT_{bottom} is weak and has little influence on the model curves on Fig. 4. These curves further show that the observations can be explained by initial excess temperatures ΔT_{bottom} in the range 100–300 °C, which is consistent with geophysical inferences²⁸ and geochemistry²⁹. Because plumes with low buoyancy fluxes cool more during upwelling, ΔT_{top} is an increasing function of B (50–70 °C for the Azores, more than 100 °C for Iceland, 200 °C for Hawaii; see below and also Supplementary Fig. 1). The estimate for Iceland is in good agreement with one obtained by comparing the observed topography along the Reykjanes ridge with that predicted by a three-dimensional convection model³⁰. The larger estimate

($\Delta T_{\text{top}} \approx 180$ °C) determined by ref. 31 using a different three-dimensional convection model relies on the amount of water present in the Icelandic mantle source, a parameter that is not well constrained.

Several studies have shown that the melting process can be strongly influenced by the presence of water in the mantle source of hot spot lavas^{7,31,32}. To take this added complexity into account, we have included the effect of water in the model (see Supplementary Information) by decreasing the initial melting rate and increasing the initial temperature of melting. Although these effects shift the model curves slightly (by at most 10–15%), the overall shape of the curves and the ranges in $(^{230}\text{Th}/^{238}\text{U})$ and $(^{231}\text{Pa}/^{235}\text{U})$ are not much affected.

The presence of water can also affect the mechanical properties of the source region where melting occurs. In particular, melting results in rapid dehydration of the source material, which can increase its viscosity by a factor of seven and thereby reduce its upwelling velocity^{7,31}. Yet, as shown by Asimow *et al.*³³, the concurrent increase in melt productivity (by a factor of 30) largely compensates for the changes due to viscosity. Taken together, the above points confirm that the observed trends of U-series activity ratios as a function of buoyancy flux (Fig. 2a and b) are mainly due to variations in upwelling velocity.

An important consequence of our modelling is that the mantle upwelling velocities can be estimated and compared with the surface plate speeds, which in turn should be comparable to upper mantle convection velocities. For plumes with low buoyancy flux (0.5 – 2 Mg s^{-1}) the calculated upwelling velocities range from 2 to 6 cm yr^{-1} , which is similar to spreading rates at mid-ocean ridges.

The width of mantle plumes

Because the analytical model of Olson *et al.*²⁴ predicts radial profiles of the vertical velocity and the excess temperature across an upwelling plume, it can be used in conjunction with the trends shown on Fig. 2 to estimate the radii of the Iceland, Azores and Hawaii plumes at the depth where melting occurs. The calculation comprises two steps.

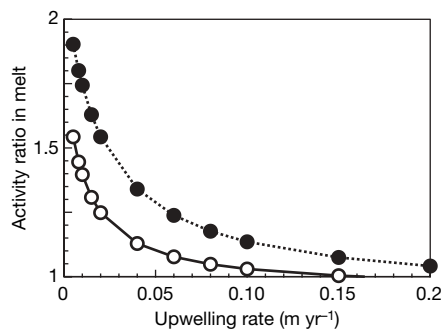


Figure 3 | Relationship between U-series activity ratios in erupted melts and mantle upwelling velocity. Open circles represent $(^{230}\text{Th}/^{238}\text{U})$ and solid circles represent $(^{231}\text{Pa}/^{235}\text{U})$. Model curves were calculated using the dynamic melting model of ref. 18 assuming constant melt productivity (that is, a linear dependence of the degree of melting on the upwelling velocity). $(^{231}\text{Pa}/^{235}\text{U})$ activity ratios are more sensitive to mantle upwelling velocity than $(^{230}\text{Th}/^{238}\text{U})$ activity ratios. The model assumes partition coefficients $D^{\text{U}} = 3.4 \times 10^{-3}$, $D^{\text{Th}} = 1.2 \times 10^{-3}$, $D^{\text{Pa}} = 7 \times 10^{-5}$, a porosity $\phi = 3 \times 10^{-3}$, and a maximum degree of melting $F_{\text{max}} = 0.05$.

First, we have used the analytical solution for axisymmetric mantle plumes derived by Olson *et al.*²⁴ to infer the temperature and vertical velocity field using the buoyancy flux and excess temperature as input parameters (see Supplementary information). Second, these parameters were used to calculate ^{230}Th - and ^{231}Pa -excess in the melt using a dynamic melting model¹⁸. The excess temperature ΔT_{top} was then adjusted to match the observed increase of ^{230}Th - and ^{231}Pa -excess as a function of distance to the centre of the hotspot (Fig. 2).

For the sake of comparison, we define the radius of the plume as the radius ($= r_{25}$) for which the upwelling velocity is 25% of its value on the plume axis. On the basis of Fig. 2, $r_{25} = 220 \pm 40$ km, 280 ± 40 km and 290 ± 50 km for Hawaii, Iceland and the Azores, respectively (see Fig. 2 legend). The radius determined for Iceland compares well with the estimate of ref. 30, based on a three-dimensional convection model. Because these radii are values at the depth of melting (60–100 km), they cannot be compared directly with the width of a plume conduit, because a plume impinging on the base of

the lithosphere broadens as it decelerates. Nevertheless, our estimates are not inconsistent with estimates from seismology^{34,35} based on inversion of body waves (at a depth of 300 km). Because the viscosity of mantle materials depends strongly on temperature, the radius of the thermal halo is significantly greater than r_{25} (the radius of the region of fast upwelling). If the mantle surrounding the plume conduit melts owing to conductive heating (for example, along a mid-ocean ridge in the case of near-ridge hotspots) and the melts rise vertically, our model predicts that the radius of the compositional and bathymetry anomalies observed around hotspots should be broader than the upwelling velocity profile depicted in Fig. 2.

In the case of the Azores, there is fairly clear evidence that the geochemical anomaly is broader than the inferred zone of extensive upwelling (see Fig. 5 and ref. 36), despite the paucity of U-series data for distances greater than about 300 km away from the plume centre. For Iceland, there is a limited data set for the Reykjanes ridge^{36,37}, which also suggests that the radiogenic isotope anomalies (Sr, Nd, Pb) extend beyond approximately 500 km, whereas the U-series signal probably tapers off at about 300–400 km away from the plume centre^{7,36}. While a more detailed examination of this issue awaits more comprehensive data sets, the available data are at least consistent with the model prediction.

The correlations between U-series disequilibria in ocean island lavas and buoyancy fluxes can be explained by coupled variations in mantle temperature and upwelling velocities. Our modelling provides evidence that hotspots are associated with increased buoyancy and temperature of the upwelling mantle: that is, they correspond to mantle plumes as defined above. The mantle upwelling velocities associated with low-buoyancy-flux hotspots are comparable to surface plate speeds, which implies that plumes are likely to be affected by upper mantle convection. It is thus possible that the position of low-buoyancy-flux plumes is not fixed.

In addition, our modelling provides constraints on the lateral extent of mantle upwelling for the Azores (220 ± 40 km), Hawaii (280 ± 40 km) and Iceland (290 ± 50 km). These are greater than some earlier estimates but comparable to estimates from seismology. The predicted temperature excesses at the top of the upwelling mantle column are 200°C (Hawaii), $70\text{--}80^\circ\text{C}$ (Iceland) and 50°C (Azores), respectively. At least in the case of the Azores and Iceland, the thermal halo surrounding the mantle plumes is wider than the region of intense upwelling. Unfortunately, our model calculations are not highly sensitive to the excess temperature in the boundary layer from which the upwelling mantle originates, making it difficult to determine the depth of this layer in the mantle.

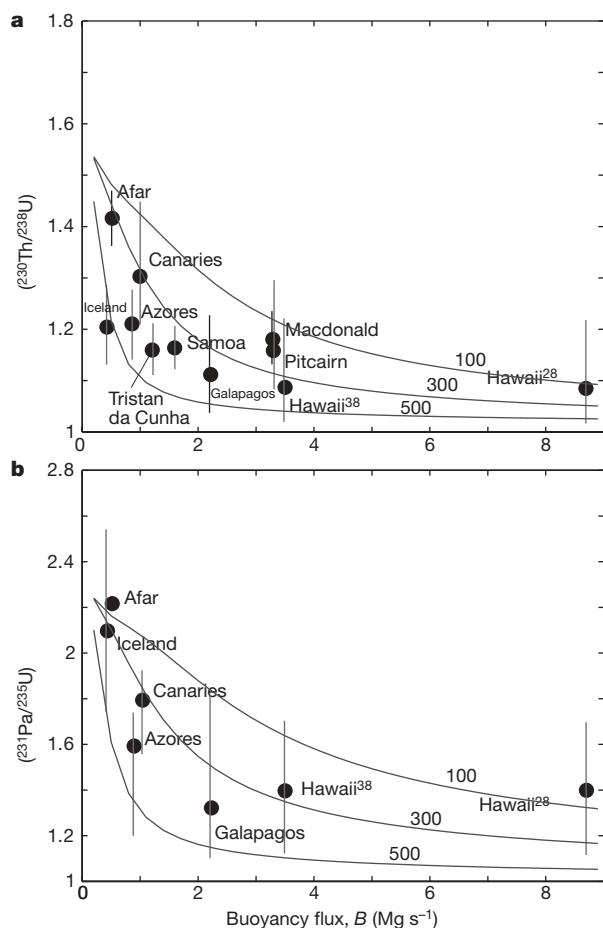


Figure 4 | Models of U-series activity ratios versus buoyancy flux for recent hotspot lavas. a, $(^{230}\text{Th}/^{238}\text{U})$. b, $(^{231}\text{Pa}/^{235}\text{U})$. U-series data sources are given in the section ‘Inferences of temperature and upwelling velocity beneath hotspots’. The plotted data represent an average of published data (data source as in Fig. 2). The vertical bars represent the dispersion of the data, that is, a 1σ standard deviation calculated based on all existing data. See caption of Fig. 1 for source of buoyancy fluxes. The model curves are labelled with the excess temperature in the boundary layer generating the plume (100, 300 and 500°C). The melting rate Γ in percentage of melt per kbar is assumed to depend on pressure³² as $\Gamma = \Gamma_0 e^{-\beta(P - P_0)}$, where β is an adjustable parameter (see also ref. 6) and P is the pressure in kbar. We assume that $\Gamma = 3$ and $0.15\% \text{ kbar}^{-1}$ at $P = 7$ and 30 kbar, respectively, similar to the values in ref. 6. Partition coefficients for U, Th and Pa are as in ref. 6.

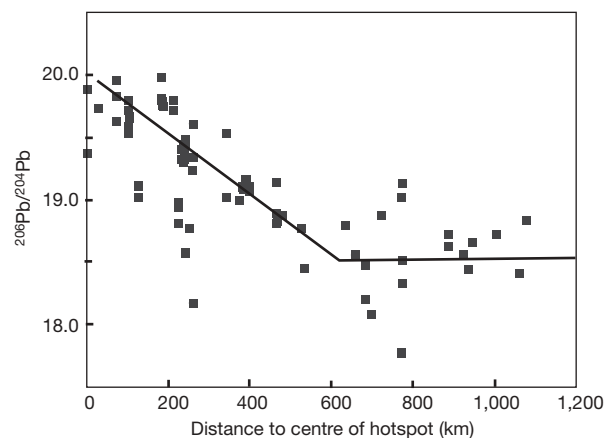


Figure 5 | $^{206}\text{Pb}/^{204}\text{Pb}$ as a function of the distance from the centre of the Azores hotspot. Although there is considerable spread in this data set, there is a marked decrease of $^{206}\text{Pb}/^{204}\text{Pb}$ for distances up to 600 km, beyond which the variation is limited. For data sources, see text.

Received 18 May; accepted 9 October 2006.

1. Foulger, G. R. *et al.* Seismic tomography shows that upwelling beneath Iceland is confined to the upper mantle. *Geophys. J. Int.* **146**, 504–530 (2001).
2. Meibom, A. *et al.* Are high $^3\text{He}/^4\text{He}$ ratios in oceanic basalts an indicator of deep-mantle plume components? *Earth Planet. Sci. Lett.* **208**, 197–204 (2003).
3. McKenzie, D. *et al.* Source enrichment processes responsible for isotopic anomalies in oceanic island basalts. *Geochim. Cosmochim. Acta* **68**, 2699–2724 (2004).
4. Bourdon, B., Joron, J. L., Claude-Ivanaj, C. & Allegre, C. J. U-Th-Pa-Ra systematics for the Grande Comore volcanics: melting processes in an upwelling plume. *Earth Planet. Sci. Lett.* **164**, 119–133 (1998).
5. Bourdon, B. & Sims, K. W. W. in *U-series Geochemistry* (eds Bourdon, B., Lundstrom, C., Henderson, G. & Turner, S. P.) 215–253 (Mineralogical Society of America, 2003).
6. Bourdon, B., Turner, S. P. & Ribe, N. M. Partial melting and upwelling rates beneath the Azores from a U-series isotope perspective. *Earth Planet. Sci. Lett.* **239**, 42–56 (2005).
7. Kokfelt, T. F., Hoernle, K. & Hauff, F. Upwelling and melting of the Iceland plume from radial variation of ^{238}U - ^{230}Th disequilibria in postglacial volcanic rocks. *Earth Planet. Sci. Lett.* **214**, 167–186 (2003).
8. Sims, K. W. W. *et al.* Porosity of the melting zone and variations in the solid mantle upwelling rates beneath Hawaii: Inferences from ^{238}U - ^{230}Th - ^{226}Ra , and ^{235}U - ^{231}Pa disequilibria. *Geochim. Cosmochim. Acta* **63**, 4119–4138 (1999).
9. Sigmarsson, O., Carn, S. & Carracedo, J. Systematics of U-series nuclides in primitive lavas from the 1730–36 eruption on Lanzarote, Canary islands, implications for the role of garnet-pyroxenite during oceanic basalt formations. *Earth Planet. Sci. Lett.* **162**, 137–151 (1998).
10. Stracke, A., Salters, V. J. M. & Sims, K. W. W. Assessing the presence of garnet-pyroxenite in the mantle sources of basalts through combined hafnium-neodymium-thorium isotope systematics. *Geochim. Geophys. Geosyst.* **1**, doi:10.1029/1999GC000013 (1999).
11. Stracke, A. *et al.* The dynamics of melting beneath Theistareykir, northern Iceland. *Geochim. Geophys. Geosyst.* **4**, 8513, doi:10.1029/2002GC000347 (2003).
12. Stracke, A., Bourdon, B. & McKenzie, D. Melt extraction in the Earth's mantle: Constraints from U-Th-Pa-Ra studies in oceanic basalts. *Earth Planet. Sci. Lett.* **244**, 97–112 (2006).
13. Bourdon, B., Langmuir, C. H. & Zindler, A. Ridge-hotspot interaction along the Mid-Atlantic Ridge between 37°30' and 40°30'N: The U-Th disequilibrium evidence. *Earth Planet. Sci. Lett.* **142**, 175–189 (1996).
14. Turner, S., Hawkesworth, C., Rogers, N. & King, P. U-Th isotope disequilibria and ocean island basalt generation in the Azores. *Chem. Geol.* **139**, 145–164 (1997).
15. Sims, K. W. W. *et al.* Mechanisms of magma generation beneath Hawaii and midocean ridges—uranium/thorium and samarium/neodymium isotopic evidence. *Science* **267**, 508–512 (1995).
16. Lundstrom, C. C., Hoernle, K. & Gill, J. B. U-series disequilibria in volcanic rocks from the Canary Islands: Plume versus lithospheric melting. *Geochim. Cosmochim. Acta* **67**, 4153–4177 (2003).
17. Vigier, N., Bourdon, B., Joron, J. L. & Allegre, C. J. U-decay series and trace element systematics in the 1978 eruption of Ardoukoba, Asal rift: timescale of magma crystallization. *Earth Planet. Sci. Lett.* **174**, 81–97 (1999).
18. Chabaux, F. & Allegre, C. J. ^{238}U - ^{230}Th - ^{226}Ra disequilibria in volcanics—a new insight into melting conditions. *Earth Planet. Sci. Lett.* **126**, 61–74 (1994).
19. MacLennan, J., McKenzie, D. & Groenvold, K. Plume-driven upwelling under central Iceland. *Earth Planet. Sci. Lett.* **2001**, 67–82 (2002).
20. DePaolo, D. J. & Stolper, E. M. Models of Hawaiian volcano growth and plume structure. Implications of results from the Hawaii Scientific Drilling Project. *J. Geophys. Res.* **101**, 11643–11654 (1996).
21. Moreira, M. *et al.* Helium and lead isotope geochemistry of the Azores Archipelago. *Earth Planet. Sci. Lett.* **169**, 189–205 (1999).
22. Richardson, C. & McKenzie, D. Radioactive disequilibria from 2D models of melt generation by plumes and ridges. *Earth Planet. Sci. Lett.* **128**, 425–437 (1994).
23. Albers, M. & Christensen, U. R. The excess temperature of plumes rising from the core-mantle boundary. *Geophys. Res. Lett.* **23**, 3567–3570 (1996).
24. Olson, P., Schubert, G. & Anderson, C. Structure of axisymmetric plumes. *J. Geophys. Res.* **98**, 6829–6844 (1993).
25. Courtillot, V., Davaille, A., Besse, J. & Stock, J. Three distinct types of hotspots in the Earth's mantle. *Earth Planet. Sci. Lett.* **205**, 295–308 (2003).
26. Boehler, R., Chopelas, A. & Zerr, A. Temperature and chemistry of the core-mantle boundary. *Chem. Geol.* **120**, 199–206 (1995).
27. Farnetani, C. G. Excess temperature of mantle plumes: The role of chemical stratification across D. *Geophys. Res. Lett.* **24**, 1583–1586 (1997).
28. Sleep, N. H. Hotspots and mantle plumes—some phenomenology. *J. Geophys. Res.* **95**, 6715–6736 (1990).
29. Schilling, J.-G. Fluxes and excess temperatures of mantle plumes inferred from their interaction with migrating mid-ocean ridges. *Nature* **352**, 397–403 (1991).
30. Ribe, N. M., Christensen, U. R. & Theissing, J. The dynamics of plume ridge interaction. 1: Ridge centered plumes. *Earth Planet. Sci. Lett.* **134**, 155–168 (1995).
31. Ito, G., Shen, Y., Hirth, G. & Wolfe, C. J. Mantle flow, melting, and dehydration of the Iceland mantle plume. *Earth Planet. Sci. Lett.* **165**, 81–96 (1999).
32. Asimow, P. D., Dixon, J. E. & Langmuir, C. H. A hydrous melting and fractionation model for mid-ocean ridge basalts: Application to the Mid-Atlantic ridge near the Azores. *Geochim. Geophys. Geosyst.* **5**, 24, doi:10.1029/2003GC000568 (2004).
33. Asimow, P. D. & Langmuir, C. H. The importance of water to oceanic mantle melting regimes. *Nature* **421**, 815–820 (2003).
34. Montelli, R. *et al.* Finite-frequency tomography reveals a variety of plumes in the mantle. *Science* **303**, 338–343 (2004).
35. Wolfe, C., Bjarnason, I. T., Van Decar, J. & Solomon, S. C. Seismic structure of the Iceland mantle plume. *Nature* **385**, 245–247 (1997).
36. Murton, B. J., Taylor, R. N. & Thirlwall, M. F. Plume-ridge interaction: a geochemical perspective from the Reykjanes ridge. *J. Petrol.* **43**, 1987–2012 (2002).
37. Peate, D. W. *et al.* ^{238}U - ^{230}Th constraints on mantle upwelling and plume-ridge interaction along the Reykjanes ridge. *Earth Planet. Sci. Lett.* **187**, 259–272 (2001).
38. Ribe, N. M. The dynamics of plume ridge interaction. 2: Off-ridge plumes. *J. Geophys. Res.* **101**, 19195–19204 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. McKenzie for the support of A.S. during his visit to IPGP. A.E.S.'s visit to Paris was supported by IPGP funds. We also thank D. McKenzie, A. Davaille, M. Moreira and C. Farnetani for numerous discussions about plumes. We are grateful to P. Olson and T. Kokfelt for comments.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.B. (bourdon@erdw.ethz.ch).

ARTICLES

The efficient interaction of indirect reciprocity and costly punishment

Bettina Rockenbach¹ & Manfred Milinski²

Human cooperation in social dilemmas challenges researchers from various disciplines. Here we combine advances in experimental economics and evolutionary biology that separately have shown that costly punishment and reputation formation, respectively, induce cooperation in social dilemmas. The mechanisms of punishment and reputation, however, substantially differ in their means for 'disciplining' non-cooperators. Direct punishment incurs salient costs for both the punisher and the punished, whereas reputation mechanisms discipline by withholding action, immediately saving costs for the 'punisher'. Consequently, costly punishment may become extinct in environments in which effective reputation building—for example, through indirect reciprocity—provides a cheaper and powerful way to sustain cooperation. Unexpectedly, as we show here, punishment is maintained when a combination with reputation building is available, however, at a low level. Costly punishment acts are markedly reduced although not simply substituted by appreciating reputation. Indeed, the remaining punishment acts are concentrated on free-riders, who are most severely punished in the combination. When given a choice, subjects even prefer a combination of reputation building with costly punishment. The interaction between punishment and reputation building boosts cooperative efficiency. Because punishment and reputation building are omnipresent interacting forces in human societies, costly punishing should appear less destructive without losing its deterring force.

Human cooperation is a paradox. Although human altruism seems unique in the animal world¹ and we are no doubt champions of reciprocity², there is a kind of social dilemma³ that comes in many guises, some of which may eventually challenge our existence. We obviously overuse public resources—for example, by over-fishing oceans, driving pension and health insurance systems to bankruptcy, and risking the collapse of the global climate through unlimited use of fossil energy. These are showpieces of the "tragedy of the commons", for which Hardin⁴ envisaged only inevitable breakdown.

However, several studies have found factors that can boost cooperation in social dilemmas, the most successful ones being costly punishment of defectors^{5–10} and reputation that is essential for success in interacting games^{11–13}. The 'public goods game' is the experimental model used to study social dilemmas¹⁴: for example, four players are given an endowment that they are free to keep for themselves or to contribute to a common pool. The content of the pool is then doubled and redistributed among the players irrespective of their contribution. If all contribute £1, everybody receives £2 (that is, a net gain of £1). If all players but one contribute, the defector has a net gain of £1.50 and the contributors £0.50 each. Individual self-interest results in defection and is at odds with group interest. Usually, initial cooperation declines quickly¹⁴, supporting Hardin's views⁴. When players are allowed to punish others costly (that is, to reduce the others' income at their own cost—for example, by investing £1 to reduce another player's income by £3), defectors are heavily punished and typically increase their contributions in future rounds, even with new partners⁷. Such altruistic punishment leads to a large increase in cooperation, but induces another type of dilemma—namely, that cooperation enhancement through severe costly punishment considerably destroys monetary gains: both the punisher and the punished lose money. Especially in the early rounds, when

punishment is heavily used, efficiency (that is, realized payoff as a fraction of maximally attainable payoff) is lower than in the absence of the punishment option. Not until cooperation has been established on a stable level is a significantly higher efficiency achieved under altruistic punishment. Thus, the group would be better off if contributions to the public good could be induced by a less destructive means¹⁵.

Various alternative mechanisms, such as symbolic punishment¹⁶ and direct rewarding of other subjects¹⁷, have failed to solve the cooperation dilemma. However, if the public goods game is incorporated in a broader context that promises rewards for those with a good reputation—that is, an 'indirect reciprocity game'—cooperation in the public goods game can be maintained at a high level¹¹. This reputation effect, 'give and you shall receive', is predicted to be rather robust¹². In indirect reciprocity situations^{2,18,19}, individuals who have helped others are given support—in other words, the supporters improve their reputation and are rewarded in turn^{20–23}. Players use withholding help in the indirect reciprocity game as a 'punishment' device for defectors in the public goods game. However, this device does not generate the dilemma of direct efficiency reduction of altruistic punishment: withholding help does not afford any monetary investment and the 'punisher' actually saves money by not rewarding a defector (who does not lose his money either; instead, he does not receive additional gains). The 'punisher' may not lose even in the long run, because an otherwise cooperative person will not destroy her image score by sometimes not giving in the indirect reciprocity game²⁴. Yet, 'punishment' through withholding help gives up the opportunity to increase efficiency through helping in the indirect reciprocity game. Thus, the accumulated evidence seems to suggest a straightforward prediction: in a social dilemma situation that allows for effective reputation building, costly punishment may

¹Department of Economics, University of Erfurt, Nordhäuser Straße 63, D-99089 Erfurt, Germany. ²Department of Evolutionary Ecology, Max-Planck-Institute of Limnology, August-Thienemann-Straße 2, D-24306 Plön, Germany.

become extinct because a much cheaper and equally powerful alternative to sustain cooperation is available. But if this is true, we face the following puzzle. Why is direct punishment of defectors a universal feature in all known human societies?

Here we try to solve this puzzle in an experiment with human volunteers. Groups of eight subjects each played 20 periods of the public goods game. We studied two treatments (Table 1). In treatment 'PUN&IR', before each period, each player can choose between joining a group in which the public goods game is followed by both costly punishing (PUN) and an indirect reciprocity (IR) game, and a group in which the public goods game is followed solely by an indirect reciprocity game; that is, we offer a choice between punishing plus indirect reciprocity and solely indirect reciprocity in treatment PUN&IR. In treatment 'PUN', before each period, each player can choose between joining a group in which the public goods game is followed by costly punishing and a group in which the public goods game is not combined with any other option; that is, we offer a choice between punishing and no additional option in treatment PUN⁹. Thus, players in PUN have the same choice as in PUN&IR but without an indirect reciprocity game in either option (see Fig. 1 and Supplementary Information). In each indirect reciprocity game, each player acts as both a 'sender' and a 'receiver'. The sender may help the receiver by sending money, which reaches the receiver in tripled amount. Before sending, the sender is informed about the receiver's 'reputation': the sender gets to know the contributions of the receiver in the actual period of the public goods game and in all previous periods, and about the previous indirect reciprocity sending decisions of the receiver. Direct reciprocity is excluded (see Methods).

Our experimental design studies a situation in which reputation cannot be avoided, but the punishment option can be chosen or avoided. We deliberately did not test whether people prefer a reputation option to a punishment option, both combined with a public goods game, because we think that this kind of choice is unlikely to occur in reality and thus humans may not be expected to have developed or evolved rules for such a decision. It seemed obvious to study whether a reputation-sensitive world can be invaded by a strategy of costly punishment because being watched can hardly be avoided in real life. We expected that players in PUN&IR would go directly for the punishment-free group because they would use the indirect reciprocity game for cooperation enhancement and avoid the detrimental effects of costly punishment^{9,15,25}. Below, we continue to use the terms 'punishment' or 'punisher' (in quotation marks) for withholding help in the indirect reciprocity game; the terms punishment or punisher are used for direct costly punishment (see refs 5–10).

The interaction is preferred and boosts efficiency

In both treatments, about 70% of the subjects initially chose the group without costly punishment (the punishment-free group or

PFG⁹). Contrary to expectations, however, subjects preferred the group with the opportunity of costly punishment (the punishment-opportunity group or POG) in the second half (periods 11–20) significantly in treatment PUN&IR ($N = 12$ groups, $z = -2.640$, $P = 0.008$, two-tailed Wilcoxon signed-ranks matched-pairs test) but not significantly in treatment PUN ($N = 6$ groups, $z = -1.787$, $P = 0.074$).

The contributions to the public goods pool were significantly higher in the POG than in the PFG both in PUN&IR ($z = -3.061$, $P = 0.002$, two-tailed Wilcoxon signed-ranks matched-pairs test) and in PUN ($z = -2.226$, $P = 0.026$; Fig. 2a). However, the contributions, either 0 or 10 money units (MUs), in the POG became significantly higher in PUN&IR than in PUN in the second half ($z = -1.976$, $P = 0.048$, two-tailed Mann–Whitney U -test). This result was achieved because in the POG the frequency of contributions of 10 MUs was higher in PUN&IR than in PUN (Fig. 2b, c; $z = -2.062$, $P = 0.039$).

The interaction reduces but concentrates punishment

The interaction between costly punishment and indirect reciprocity not only increased the contributions to the public good but also significantly reduced the number of punishment points allocated per member of the POG in PUN&IR as compared with PUN (Fig. 3a; $z = -2.767$, $P = 0.006$, two-tailed Mann–Whitney U -test). Nevertheless, the punished free-riders were more heavily punished in PUN&IR than in PUN (Fig. 3b; $z = -2.313$, $P = 0.018$). Thus, although fewer group members were punished, punishment of a free-rider, that is a non-cooperator, was even more severe in the presence of the indirect reciprocity game than in its absence. This

Table 1 | Experimental design

Endogenous choice	Experimental treatments (exogenous)			
	PUN&IR		PUN	
	Indirect reciprocity exists (reputation building possible)		Indirect reciprocity ruled out (reputation building impossible)	
POG Public good with punishment opportunity	PUN&IR	POG	PUN	POG
PFG Public good without punishment opportunity	PUN&IR	PFG	PUN	PFG

PFG, punishment free group; POG, punishment opportunity group; PUN, public good game with costly punishment opportunity; PUN&IR, public good game with costly punishment opportunity and indirect reciprocity game.

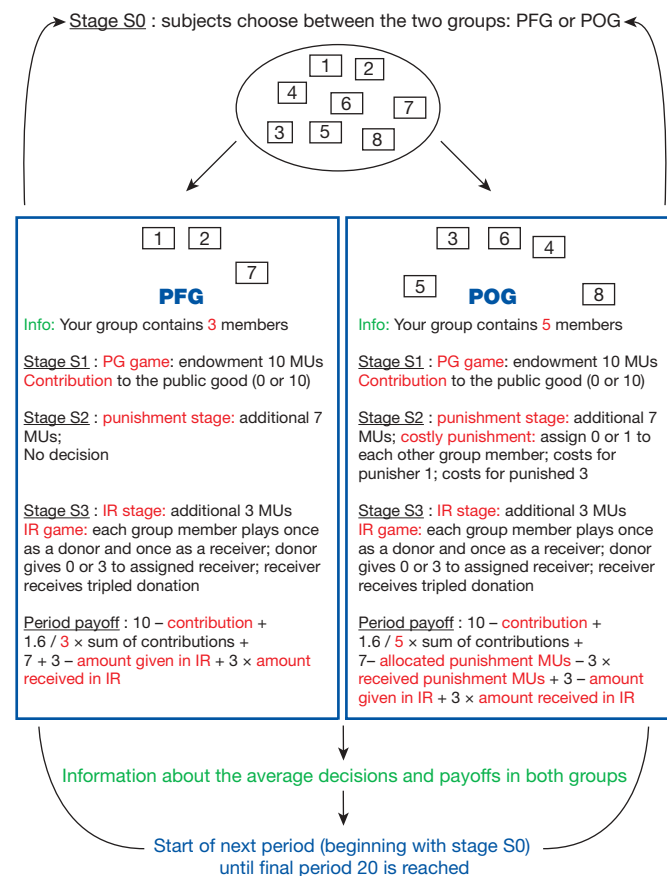


Figure 1 | The course of a period. Shown is the course of a period in treatment PUN&IR with the stages S0, S1, S2 and S3. In treatment PUN, the stages S0, S1 and S2 are played as shown and the period ends after the termination of stage S2. In all treatments, each of the 20 periods has an identical structure. PG game, public goods game.

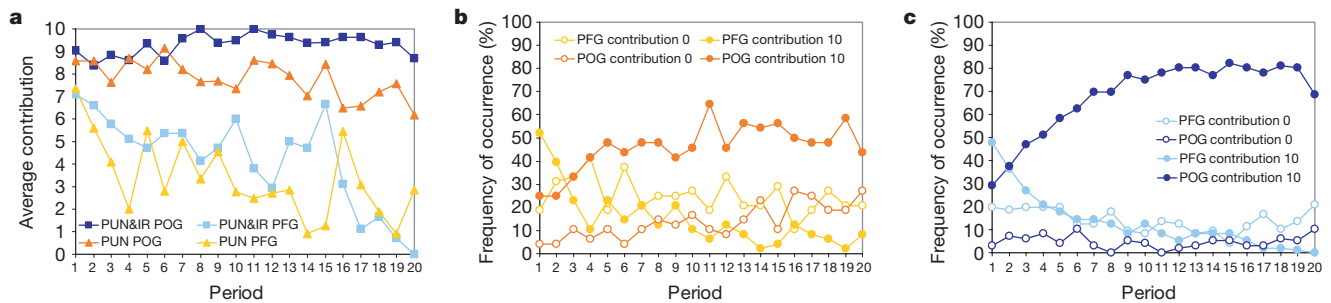


Figure 2 | Contributions to the public good. **a**, Contributions to the public good are higher in the POG than in the PFG in both treatments. **b**, In treatment PUN, subjects contributing 10 MUs become a relative, but never

is notable, because it might have been expected that the costly punishment acts would be avoided completely at this stage and that punishment would be moved (denoted 'punishment') to the indirect reciprocity stage. Contributions to the public goods pool per allocated punishing point are about three times higher in PUN&IR than in PUN (Fig. 3c; $z = -2.363$, $P = 0.015$). Thus, subjects could generate higher contributions with fewer punishment expenses (Fig. 3d; $z = -2.160$, $P = 0.030$ in the second half).

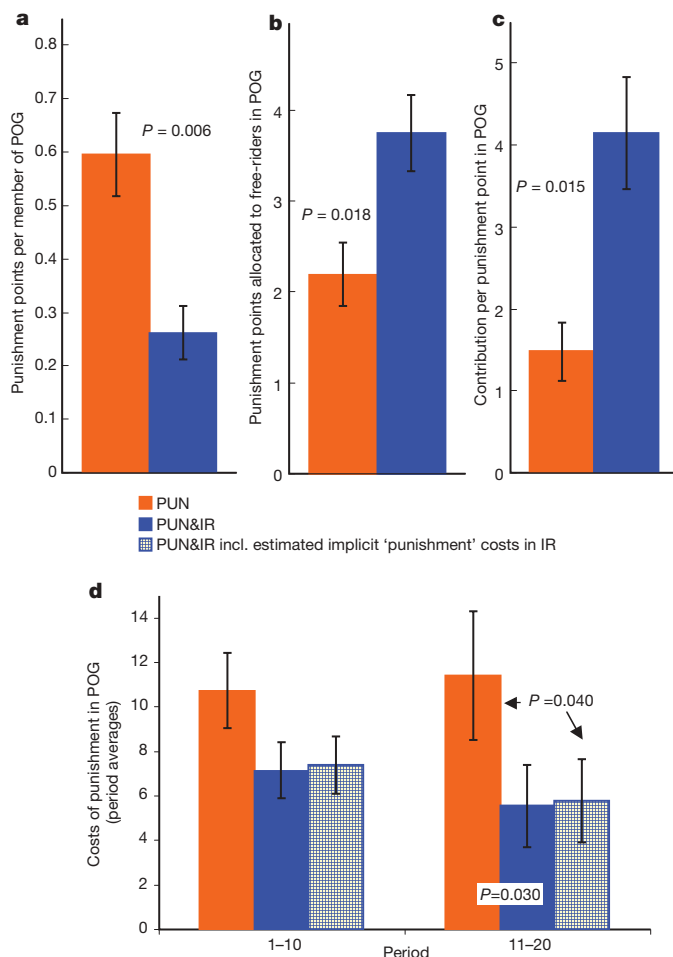


Figure 3 | Punishment behaviour. **a**, The punishment points allocated per member are significantly higher in PUN than in PUN&IR. **b**, Free-riders are significantly more severely punished in PUN&IR than in PUN. **c**, The contribution achieved per punishment point is significantly higher in PUN&IR than in PUN. **d**, The punishment costs are higher in PUN than in PUN&IR, even if the estimated implicit costs of 'punishment' in the indirect reciprocity stage are added, especially in the second half (see text for statistics). Data are the mean $\pm$ s.e.m. per group.

an absolute majority. **c**, In treatment PUN&IR, subjects contributing 10 MUs quickly become an absolute majority. Data are the mean per group for each period.

In addition, when the costs of punishment of both the punisher and the punished were added to the profits from the public goods game, those who contributed 10 MUs to the public pool in the POG came very close to the social optimum—that is, full contribution of all subjects and no punishment acts (Fig. 4a, horizontal line at 27)—and had a higher net payoff in PUN&IR than in PUN ($z = -2.435$, $P = 0.015$; two-tailed Mann–Whitney U -test). Those who contributed 0 MUs were not significantly better off in PUN&IR than in PUN ($z = -1.032$, $P = 0.302$). In both treatments, their payoff seems to come close to the Nash equilibrium of all players defecting and no punishment acts executed (Fig. 4a, horizontal line at 17). Even if we compare the efficiency in the public goods game including all costs of punishment of the combination of both options (with and without punishment) of a treatment, we find that PUN&IR is more efficient

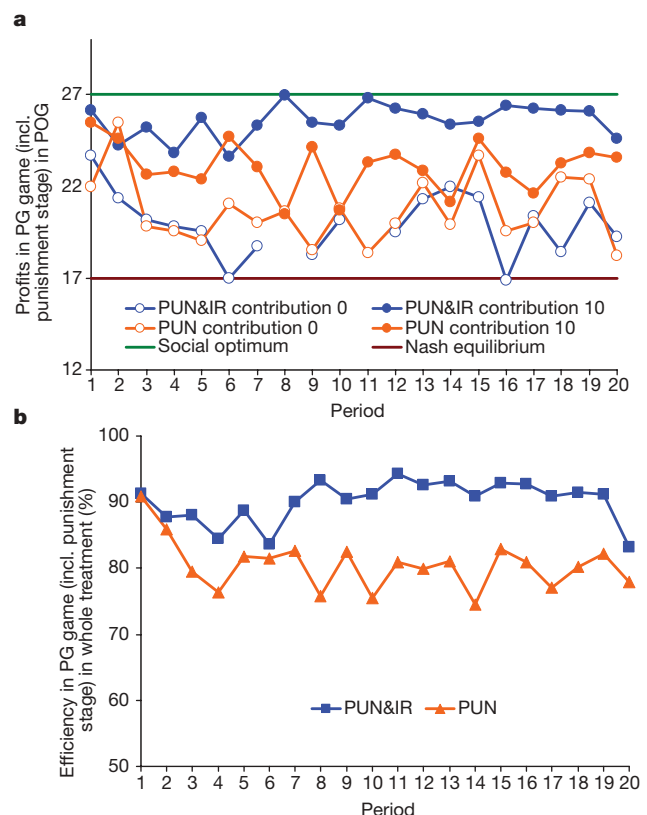


Figure 4 | Payoffs in the public goods game including the punishment stage and efficiency. **a**, Payoffs are highest for the full contributors in the POG in PUN&IR. Horizontal line at 27 indicates the social optimum; horizontal line at 17 indicates the Nash equilibrium. **b**, Treatment efficiency, that is, the efficiency of the whole treatment combining both groups (PFG and POG), is higher in PUN&IR than in PUN. Data are the mean per group.

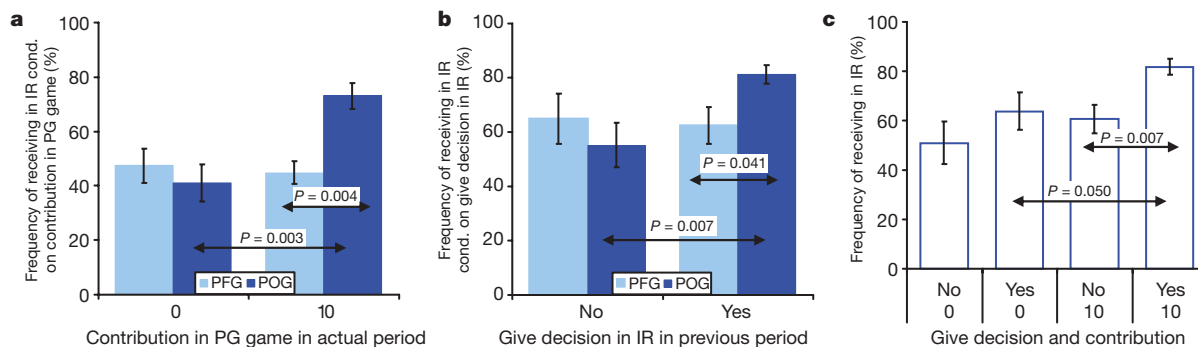


Figure 5 | Receiving in the indirect reciprocity game. **a**, High contributors in the public goods game receive significantly more often in the indirect reciprocity (IR) game in the POG. **b**, Subjects who gave in the previous period of the indirect reciprocity game receive significantly more often in the actual period than subjects who did not give in the previous period. **c**, High

contributors in the public goods game who gave in the previous period of the indirect reciprocity game receive significantly more often than high contributors in the public goods game who did not give in the previous period of the indirect reciprocity game. Data are the mean $\pm$ s.e.m. per group (see text for statistics).

than PUN (Fig. 4b; $z = -2.557$, $P = 0.010$, two-tailed Mann–Whitney U -test).

These results suggest that the presence of the indirect reciprocity game has a positive influence on the cooperation behaviour in the public goods game. Indeed, we find an interaction among indirect reciprocity, costly punishment and contributions to the public pool. Those subjects who had contributed 10 MUs to the public pool received significantly more often in the indirect reciprocity game than those who had contributed nothing (Fig. 5a; $z = -2.934$, $P = 0.003$, two-tailed Wilcoxon signed-ranks matched-pairs test) in the POG but not in the PFG ($z = -0.236$, $P = 0.814$). In addition, contributing 10 MUs was more often ‘rewarded’ in the POG than in the PFG ($z = -2.845$, $P = 0.004$). Thus, defectors in the POG are hit twice: they are punished once in the punishment stage and ‘punished’ a second time in the indirect reciprocity game.

Implicit costs of ‘punishment’ in the indirect reciprocity game

In addition to the calculated explicit costs of punishment in the punishment stage, the ‘punishment’ in the indirect reciprocity game induces implicit ‘punishment’ costs by giving up the opportunity to increase the efficiency by helping. The implicit ‘punishment’ costs can be estimated from the difference in the likelihood of receiving between subjects who contributed to the public good and those who did not, under the condition that they had the same giving behaviour in the previous period of the indirect reciprocity game. For example, a contributor who gave in the previous period of the indirect reciprocity game receives with a likelihood of 85.3%, whereas a free-rider who gave in the previous period of the indirect reciprocity game receives with a likelihood of 59.1%. The assumption that the difference in the likelihood of receiving between a contributor and a free-rider is attributed solely to the ‘punishment’ of free-riders in the indirect reciprocity game provides an upper bound for the implicit ‘punishment’ costs in the indirect reciprocity game. The estimation of these implicit ‘punishment’ costs shows that they are extremely low (on average, 0.30 per period in the first ten periods and 0.23 per period in the last ten periods). Thus, adding these implicit ‘punishment’ costs to the explicit punishment costs (Fig. 3d) does not change the picture that the interaction of costly punishment and reputation building through indirect reciprocity increases efficiency (the punishment costs in PUN are higher than the sum of the explicit punishment and the implicit ‘punishment’ costs in PUN&IR: $z = -2.064$, $P = 0.040$).

Reputation building is effective in the interaction

The indirect reciprocity game ‘worked’ by itself only in the POG. Individuals who had helped others in the previous indirect reciprocity game were significantly more likely to receive money in the current indirect reciprocity game (Fig. 5b) only in the POG

($z = -2.714$, $P = 0.007$, in PFG: $z = -0.561$, $P = 0.574$, two-tailed Wilcoxon signed-ranks matched-pairs test). Thus, giving in the indirect reciprocity game was more often rewarded in the POG than in the PFG ($z = -2.045$, $P = 0.041$). The efficiency (that is, the percentage of the maximal attainable payoff actually achieved) of indirect reciprocity was significantly higher in the POG than in the PFG ($z = -2.805$, $P = 0.005$). These results suggest that the subjects who chose the POG cared more about other people’s reputation than those who avoided the POG. It would be interesting to study the motivation of the few subjects who finally decided to avoid the POG in PUN&IR. These subjects seem to be pure egoists who prefer not to have a direct punishment opportunity and do not discriminate in their helping decisions between those with a low or a high reputation or between those who cooperated and those who defected in the public goods game. Nevertheless, in both groups the ‘good guys’ in the public goods game were likely to be ‘donors’ in the indirect reciprocity game (POG: $\rho = 0.738$, $P = 0.01$; PFG: $\rho = 0.887$, $P = 0.0001$, Spearman rank correlation).

Those who contributed to the public pool thus seem more at risk to damage their reputation by not giving in indirect reciprocity rounds than those who defected in the public goods game. Giving or not giving in the indirect reciprocity game was significantly taken into account in the next indirect reciprocity game when both kinds of receivers had contributed 10 MUs to the public pool (Fig. 5c; $z = -2.707$, $P = 0.007$, two-tailed Wilcoxon signed-ranks matched-pairs test) but not when receivers had not contributed ($z = -1.472$, $P = 0.141$).

Free-riders are disciplined more in the interaction

We indicated above that the ‘double punishment’ in PUN&IR increased the costs of defecting in the public goods game. To study the consequences for individual subjects, we investigated the behaviour of the two least cooperative subjects of each group, that is, those who contributed least and second least to the public goods pool during the 20 periods. On average, these two defected in 7.4 ± 1.2 periods (mean $\pm$ s.e.m.) in PUN&IR and in 12.3 ± 1.9 periods in PUN ($z = -2.255$, $P = 0.024$, two-tailed Mann–Whitney U -test). Notably, the number of periods in which the two least cooperative subjects chose the POG was not significantly different between PUN&IR (12.8 ± 1.5 times) and PUN (11.0 ± 1.6 times, $z = -1.176$, $P = 0.240$). However, they defected significantly less often when in the POG in PUN&IR (2.6 ± 0.8 times) than in the POG in PUN (5.3 ± 0.7 times, $z = -2.674$, $P = 0.007$). Thus, the least cooperative subjects were more cooperative in PUN&IR than in PUN, and they did not avoid the POG in either treatment.

Discussion and general implications

The interaction between the potential for costly punishment and building personal reputation in public goods games produced

unexpected results. Although subjects could choose a punishment-free option where reputation building alone would allow for high levels of cooperation in a public goods game^{11,12}, the great majority chose the punishment option that was also combined with reputation building. It is worthwhile to mention that most subjects preferred the group with the punishment option even though the private benefits from one unit contribution to the public good are diluted by the larger number of group members. However, the interaction with indirect reciprocity reduced punishment behaviour to almost a third and thus largely removed the efficiency dilemma of costly punishment. Indirect reciprocity did not simply replace punishment. We observed not only a net increase in contributions to the public goods pool as compared with a control (PUN), but also a boost in efficiency. This marked effect was achieved by concentrating the remaining punishing acts on the smaller number of free-riders in PUN&IR, which were thus more heavily punished than the free-riders in PUN. Moreover, the few free-riders in PUN&IR were hurt a second time through withheld help in the indirect reciprocity game, which amplified the punishment effect. These strong incentives to cooperate probably induced not only the majority of subjects but even the 'worst' free-riders per group to be more cooperative when punishment was combined with reputation building.

No theoretical prediction from a mathematical model capturing the specific situation of our study has been available; our unexpected results might thus be an incentive for theorists to study the interaction between punishment and reputation building formally. Such an analysis may also provide deeper insights into the sensitivity of the interaction between punishment and reputation building concerning, for example, group size, group composition, completeness of information about others, or impact of punishment power.

Punishment seems to have a solid foundation in human interactions. Using and reacting to costly punishment has an emotional representation in the brain^{26,27} and thus seems to be a 'hardwired' prerequisite of our social life. In addition, the finding that we take sophisticated signs of 'others may watch' into account^{28,29} suggests that 'don't lose your reputation'³⁰ is an unconscious imperative that guides social behaviour. Punishment and reputation do not seem to be substitutes that may easily replace one another, but omnipresent interacting mechanisms. Costly punishment can evolve because it purchases the reputation of someone who demands cooperation in the public goods game³¹. Of note, overwhelmingly social and fair outcomes are achieved in a model when punishment and reputation in groups of three players in spatial public goods games on a lattice are combined³². Reputation mechanisms generate an environment in which the execution of punishment has to be less frequent, however, without taking away its deterring force. The interaction of both mechanisms not only comes closer to real life but also provides a convenient, low-cost solution to social dilemmas, keeping the teeth of costly punishment covered but ready to bite. The attenuation of the destructive force of punishment through reputation may help to explain its evolutionary survival. The study of the interaction of two obviously unavoidable mechanisms, punishment and reputation, for maintaining human cooperation will certainly help to disentangle more of our complex social and unsocial nature in the future.

METHODS

Subjects. A total of 144 undergraduate students from the University of Erfurt voluntarily participated in 18 experimental sessions with eight subjects each.

Game. Each of the 20 experimental periods (Fig. 1) proceeded with a group choice stage (S0), a voluntary contribution stage (S1), a punishment opportunity stage (S2), and an indirect reciprocity stage (S3). In treatment PUN&IR, all four stages were played in 12 independent sessions; in treatment PUN, the stages S0, S1 and S2 were played in 6 independent sessions.

In stage S0, the subjects simultaneously and independently choose between a POG and a PFG in which punishment is not possible. Each subject is informed about the number of subjects in each group. The remaining period interactions take place solely within the group chosen in S0. In stage S1, the public goods game, each subject is endowed with 10 MUs and may contribute either 0 or

10 MUs. The contributions are multiplied by 1.6 to create a benefit for the whole group that is equally distributed among the group members; in other words, if a group consists of n members, each member profits by $1.6/n$ MUs from each MU contributed. The MUs not contributed are transferred to the subject's private account.

At the beginning of stage S2, each subject receives an additional 7 MUs independent of the affiliation choice in S0. In the PFG, these MUs are directly transferred to the subject's private account without any decisions required. In the POG, these MUs may be used to punish other members of the POG. Punishing costs the punisher 1 MU and the punished subject 3 MUs. The MUs not used for punishment are transferred to the subject's private account.

In stage S3, each subject receives an additional 3 MUs independent of the affiliation choice in S0. These MUs may be sent to another member of the same group, the 'receiver'. The MUs are tripled when sent, such that the receiver achieves 9 MUs. If the sender does not send, the sender keeps the 3 MUs and nothing reaches the receiver. The sender is informed both about the contributions of the receiver in the actual period of the public goods game and in all previous periods, and about the previous indirect reciprocity sending decisions of the receiver in addition to the contributions to the public good of the past receiver's receiver. Each subject acts as both a sender and a receiver in each period. To exclude direct reciprocity, the matching between the sender and the receiver ensures that the receiver has not previously been the sender's sender and will not be in the future.

At the end of the period each participant receives anonymous summary information about both groups. At the beginning of the experiment, subjects received written and oral instructions. At the end of the experiment, subjects privately received their experimental earnings in cash. All experimental decisions were made on a computer screen using the experimental software Z-Tree³³.

Received 22 June 2006; accepted 31 August 2006.

1. Fehr, E. & Fischbacher, U. The nature of human altruism. *Nature* **425**, 785–791 (2003).
2. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity. *Nature* **437**, 1291–1298 (2005).
3. Dawes, R. Social dilemmas. *Annu. Rev. Psychol.* **31**, 169–193 (1980).
4. Hardin, G. The tragedy of the commons. *Science* **162**, 1243–1248 (1968).
5. Boyd, R. & Richerson, P. J. Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethol. Sociobiol.* **13**, 171–195 (1992).
6. Fehr, E. & Gächter, S. Cooperation and punishment in public goods experiments. *Am. Econ. Rev.* **90**, 980–994 (2000).
7. Fehr, E. & Gächter, S. Altruistic punishment in humans. *Nature* **415**, 137–140 (2002).
8. Fehr, E. & Rockenbach, B. Human altruism: economic, neural, and evolutionary perspectives. *Curr. Opin. Neurobiol.* **14**, 784–790 (2004).
9. Gürrer, Ö., Irlenbusch, B. & Rockenbach, B. The competitive advantage of sanctioning institutions. *Science* **312**, 108–111 (2006).
10. Sutter, M., Haigner, S. & Kocher, M. Choosing the stick or the carrot? Endogenous institutional choice in social dilemma situations. CEPR Discussion Paper no. 5497 (Centre for Economic Policy Research, London, 2006).
11. Milinski, M., Semmann, D. & Krambeck, H.-J. Reputation helps solve the 'tragedy of the commons'. *Nature* **415**, 424–426 (2002).
12. Panchanathan, K. & Boyd, R. Indirect reciprocity can stabilize cooperation without the second-order free rider problem. *Nature* **432**, 499–502 (2004).
13. Milinski, M., Semmann, D., Krambeck, H.-J. & Marotzke, J. Stabilizing the Earth's climate is not a losing game: supporting evidence from public goods experiments. *Proc. Natl Acad. Sci. USA* **103**, 3994–3998 (2006).
14. Ledyard, J. O. in *Handbook of Experimental Economics* (eds Kagel, J. H. & Roth, A. E.) 111–194 (Princeton Univ. Press, Princeton, 1995).
15. Botelho, A., Harrison, G. W., Pinto, L. M. C. & Rutström, E. E. Social norms and social choice. NIMA Working Paper no. 30 (Núcleo de Investigação em Microeconomia Aplicada, Univ. Minho, Braga, Portugal, 2005).
16. Masclet, D., Noussair, C., Tucker, S. & Villeval, M.-C. Monetary and nonmonetary punishment in the voluntary contributions mechanism. *Am. Econ. Rev.* **93**, 366–380 (2003).
17. Sefton, M., Shupp, R. & Walker, J. The effect of sanctions in provision of public goods. CeDex Discussion Paper no. 2002-02 (The Centre for Decision Research and Experimental Economics, Nottingham, 2002).
18. Alexander, R. D. *The Biology of Moral Systems* (de Gruyter, New York, 1987).
19. Nowak, M. A. & Sigmund, K. Evolution of indirect reciprocity by image scoring. *Nature* **393**, 573–577 (1998).
20. Wedekind, C. & Milinski, M. Cooperation through image scoring in humans. *Science* **288**, 850–852 (2000).
21. Bolton, G. E., Katok, E. & Ockenfels, A. Cooperation among strangers with limited information about reputation. *J. Public Econ.* **89**, 1457–1468 (2005).
22. Seinen, I. & Schram, A. Social status and group norms: Indirect reciprocity in a repeated helping experiment. *Eur. Econ. Rev.* **50**, 581–602 (2006).
23. Bshary, R. & Grutter, A. S. Image scoring and cooperation in a cleaner fish mutualism. *Nature* **441**, 975–978 (2006).

24. Milinski, M. in *Cooperation in Primates and Humans. Mechanisms and Evolution* (eds Kappeler, P. M. & van Schaik, C. P.) 265–278 (Springer, New York, 2005).
25. Fehr, E. & Rockenbach, B. Detrimental effects of sanctioning on human altruism. *Nature* **422**, 137–140 (2003).
26. De Quervain, D. J.-F. et al. The neural basis of altruistic punishment. *Science* **305**, 1254–1258 (2004).
27. Singer, T. et al. Empathic neural responses are modulated by the perceived fairness of others. *Nature* **439**, 466–469 (2006).
28. Haley, K. J. & Fessler, D. M. T. Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evol. Hum. Behav.* **26**, 245–256 (2005).
29. Bateson, M., Nettle, D. & Roberts, G. Cues of being watched enhance cooperation in a real-world setting. *Biol. Lett.* **2**, 412–414 (2006).
30. Fehr, E. Human behaviour: Don't lose your reputation. *Nature* **432**, 449–450 (2004).
31. Sigmund, K., Hauert, C. & Nowak, M. A. Reward and punishment. *Proc. Natl Acad. Sci. USA* **98**, 10757–10762 (2001).
32. Brandt, H., Hauert, C. & Sigmund, K. Punishment and reputation in spatial public goods games. *Proc. R. Soc. Lond. B* **270**, 1099–1104 (2003).
33. Fischbacher, U. Z-tree. Zurich Toolbox for Readymade Economic Experiments—Experimenter's Manual. Working Paper no. 21 (Institute for Empirical Research in Economics, Univ. Zurich, 1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank S. Große, Ö. Gürerk, R. Sommerfeld and H.-J. Krambeck for support, and the University of Erfurt and the Max Planck Society for financial funding.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to B.R. (bettina.rockenbach@uni-erfurt.de) or M.M. (milinski@mpil-ploen.mpg.de).

ARTICLES

Thymic selection threshold defined by compartmentalization of Ras/MAPK signalling

Mark A. Daniels¹, Emma Teixeira¹, Jason Gill², Barbara Hausmann¹, Dominique Roubaty¹, Kaisa Holmberg³, Guy Werlen⁴, Georg A. Holländer², Nicholas R. J. Gascoigne³ & Ed Palmer¹

A healthy individual can mount an immune response to exogenous pathogens while avoiding an autoimmune attack on normal tissues. The ability to distinguish between self and non-self is called 'immunological tolerance' and, for T lymphocytes, involves the generation of a diverse pool of functional T cells through positive selection and the removal of overtly self-reactive thymocytes by negative selection during T-cell ontogeny. To elucidate how thymocytes arrive at these cell fate decisions, here we have identified ligands that define an extremely narrow gap spanning the threshold that distinguishes positive from negative selection. We show that, at the selection threshold, a small increase in ligand affinity for the T-cell antigen receptor leads to a marked change in the activation and subcellular localization of Ras and mitogen-activated protein kinase (MAPK) signalling intermediates and the induction of negative selection. The ability to compartmentalize signalling molecules differentially in the cell endows the thymocyte with the ability to convert a small change in analogue input (affinity) into a digital output (positive versus negative selection) and provides the basis for establishing central tolerance.

Thymocyte selection occurs through interactions of the T-cell antigen receptor (TCR) with specific complexes of self-peptide and the major histocompatibility complex (pMHC) expressed on thymic antigen-presenting cells (APCs). Weak TCR–pMHC interactions do not support thymocyte survival (death by neglect); strong interactions lead to thymocyte apoptosis, lineage deviation or receptor editing (collectively called negative selection); and interactions between these extremes lead to the development of mature T cells (positive selection)¹. The TCR is an unusual receptor that can signal different cell fates, and two models have been proposed to describe how this differential signalling occurs². The avidity model suggests that the level of TCR occupancy determines selection outcome. This model arose from studies in which strong agonist ligands administered at extremely low doses lead to positive selection^{3,4}. In many cases, however, the lymphocytes generated in this way develop into CD4-expressing (CD4⁺) regulatory T cells^{5,6} or represent a distinct lineage of CD8 $\alpha\alpha$ ⁺ T cells with regulatory properties^{7–9}. The affinity model proposes that selection outcome is established by the affinity of the TCR for a pMHC complex. In support of this, kinetic proofreading predicts that the length of TCR engagement determines positive and negative selecting signals^{10,11}. The importance of ligand affinity and half-life are well documented^{12–15}. TCR–pMHC affinities measured by surface plasmon resonance (SPR) illustrate the correlation among ligand affinity, half-life and selection outcome; however, SPR experiments have been unable to resolve the contribution of the co-receptor^{16,17}, which is known to be crucially important for the selection decision¹⁸. In addition, these assays cannot assess the contribution of TCR extrinsic factors necessary for the translation of ligand engagement such as glycosylation state¹⁹ or membrane microenvironment²⁰.

A longstanding issue remains concerning how the TCR reads the parameters of ligand engagement and signals these distinct cell fates. The role of MAPKs in selection has been well studied²¹. Phosphorylation of extracellular-signal-regulated kinase (ERK) is essential for positive selection but its role in negative selection is less clear^{21–25}. Negative selection is dependent on activation of Jun amino-terminal kinase (JNK)²⁶; however, positive and negative selectors can activate both pathways. Therefore, it is not clear how activation of ERK and JNK establishes selection outcome unless another level of regulation controls their function.

We have used the OT-I TCR to address these issues. We identified variants of the agonist peptide that define the boundary between positive and negative selection. Although ligands lying close to the selection boundary bind the TCR with only slightly different affinities, positive and negative selecting ligands organize early TCR signalling intermediates into distinct subcellular compartments. These findings elucidate part of the mechanism used by thymocytes to interpret selecting ligands and have implications for central tolerance and autoimmunity.

Defining the thymic selection threshold

The OT-I TCR recognizes the chicken ovalbumin peptide, SIINFEKL (OVA), presented by K^b as an agonist. We generated OVA variants to establish a hierarchy of ligands with a wide range of potencies and to narrow the gap between positive and negative selecting ligands described for this system²⁷. Expression of CD69 is considered to be a reflection of the capacity of a ligand to induce TCR signals independent of selection outcome²⁸. Therefore, peptides were tested for their ability to induce CD69 expression on pre-selection CD4⁺CD8⁺

¹Laboratory of Transplantation Immunology and Nephrology, Department of Research, University Hospital–Basel, Hebelstrasse 20, 4031 Basel, Switzerland. ²Pediatric Immunology, Center for Biomedicine and The University Children's Hospital of Basel, Mattenstrasse 28, 4058 Basel, Switzerland. ³Department of Immunology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA. ⁴Department of Cell Biology & Neuroscience, Rutgers, The State University of New Jersey, 604 Allison Road, Piscataway, New Jersey 08854-8082, USA.

double-positive thymocytes from OT-I transgenic $Rag^{-/-}\beta 2m^{-/-}$ mice (hereafter called 'pre-selection OT-I double-positive thymocytes'). Potencies were calculated by correcting the effector concentration for half-maximum response (EC_{50}) values for small differences in peptide affinity for K^b and were normalized to OVA (Fig. 1a and Supplementary Fig. 1).

The selection potential of ligands was tested in fetal thymic organ culture (FTOC). OVA, Q4 and Q4R7 induced a loss of double-positive thymocytes, generated only limited numbers of $CD4^+CD8\beta^+$ ($CD8\alpha\beta$ single-positive) thymocytes and were classified as negative selectors. In contrast, Q4H7, Q7, G4 and E1 induced positive selection, resulting in substantial numbers of $CD8\alpha\beta$ single-positive thymocytes²⁹ (Fig. 1b). By plotting $1/(\text{ligand potency})$ versus the production of $CD8\alpha\beta$ single-positive cells (Fig. 1c), an abrupt transition from positive to negative selection was observed. The weakest negative selector, Q4R7, was only twofold more potent than the strongest positive selector, Q4H7. Importantly, altering peptide dose did not change the selection outcome in FTOC. Q4R7 remained a negative selector and Q4H7 remained a positive selector over a broad range of concentrations (data not shown). T4 was the only peptide that showed variation in selection outcome on the basis of peptide dose (Fig. 1b). The potency of the T4 peptide was between Q4R7 and Q4H7. In addition, only negative selecting ligands ($OVA > Q4 > Q4R7$) induced a Ca^{2+} flux or cytolysis

by mature peripheral T cells (Supplementary Fig. 1). Despite its ability to induce negative selection at high doses in FTOC, T4 was a poor antigen for peripheral T cells. These data imply that the selection border defined in FTOC is relevant for peripheral T cells and demonstrate the sharpness of the threshold.

To estimate the TCR affinity of the ligands comprising the selection boundary, we measured tetramer binding; which correlates with monomeric TCR-pMHC affinities^{30,31}, is performed on live cells and involves the participation of CD8 (refs 32, 33). The binding characteristics of tetramers were determined on pre-selection OT-I double-positive thymocytes at 37 °C (ref. 33). The dissociation constant (K_d) was calculated by nonlinear regression analysis and confirmed by homologous competition experiments (data not shown). The tetramer binding curves for Q4R7 (weakest negative selector), T4 (border ligand) and Q4H7 (strongest positive selector) overlapped (Fig. 2a). Their K_d values (Q4R7, 48 ± 9.5 nM; T4, 55 ± 10.1 nM; Q4H7, 51 ± 9.1 nM; $n = 7$, $P = 0.455$) and their half-lives ($t_{1/2}$) were not significantly different (Table 1). However, heterologous competition assays showed that Q4R7 was more efficient than Q4H7 at inhibiting the binding of OVA tetramers (data not shown).

A tetramer can bind to a thymocyte through CD8 alone¹⁹, the TCR alone, or the TCR and CD8. We therefore considered the possibility that, despite quantitative similarities, there may be qualitative differences in how each ligand is bound to the thymocyte. To dissect these differences, tetramer binding in the absence of CD8 participation was quantified by using K^b tetramers carrying the D227K mutation (hereafter termed the K^b mutant), which precludes CD8 binding³⁴. Without CD8 participation, binding differences between Q4R7 and Q4H7 were readily seen (Q4R7 K^b mutant, 282 ± 49 nM; Q4H7 K^b mutant, 721 ± 51 nM; $n = 4$, $P \leq 0.02$). The K_d for the T4 K^b mutant tetramer (576 ± 97 nM) lay between that of the Q4R7 and that of Q4H7 K^b mutant tetramers. We determined binding to CD8 on pre-selection double-positive thymocytes using wild-type K^b tetramers loaded with the null peptide VSV (Fig. 2b and Table 1). The K_d of K^b binding to only CD8 (490 ± 89 nM) was similar to the K_d of T4 K^b mutant tetramers binding to only the OT-I TCR (576 ± 97 nM; $n = 4$, $P = 0.29$). Therefore, at the boundary between positive and negative selection, the pMHC affinity for the TCR was roughly equal to its affinity for CD8.

To evaluate the individual contributions of the TCR and CD8 to ligand binding, tetramer dissociation assays were done in the

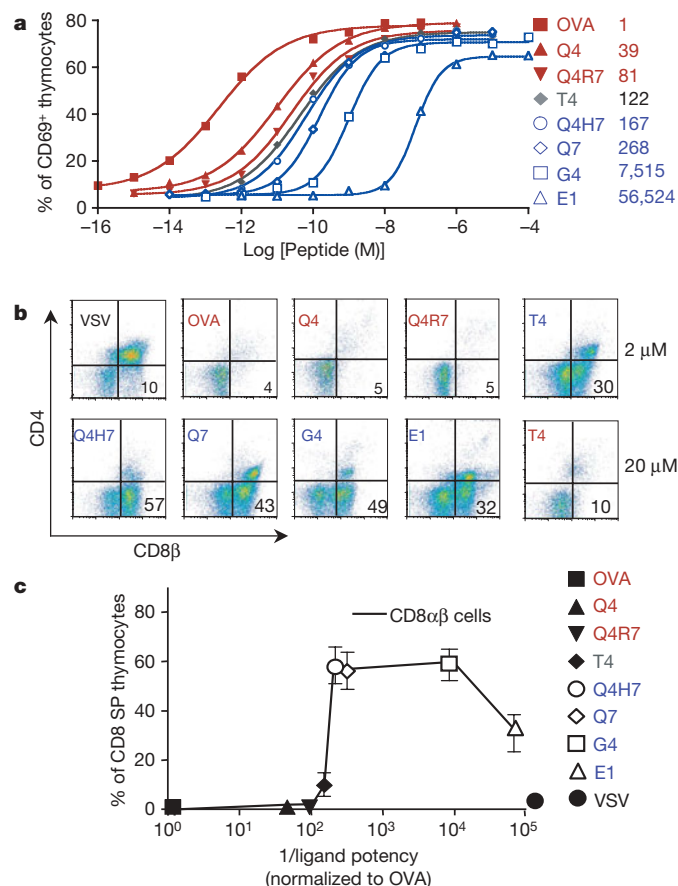


Figure 1 | Defining ligand potency and the thymic selection boundary. **a**, OT-I pre-selection thymocytes were stimulated with peptide-pulsed T2 K^b -expressing APCs and the CD69 response was determined. Numbers on the right represent $1/(\text{ligand potency})$ normalized to OVA. **b**, FTOCs were performed to establish the ability of a peptide to induce positive or negative selection (presence or absence of $CD8\alpha\beta$ single-positive cells, respectively). Shown are the CD4/CD8 β expression profiles for peptide doses of 2 μ M (top) and 20 μ M (bottom). The number in each dot plot is the percentage of $CD8\alpha\beta$ single-positive thymocytes. **c**, Percentage of $CD8\alpha\beta$ single-positive (SP) thymocytes generated in FTOC as a function of normalized ligand potency for all peptides at a dose of 20 μ M. Error bars are s.d.

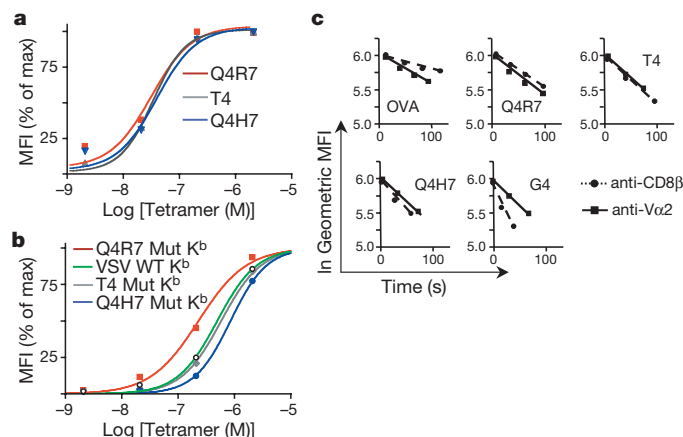


Figure 2 | Ligand affinity and contributions of CD8 and TCR to ligand binding. **a**, **b**, OT-I double-positive thymocytes were incubated at 37 °C with K^b (WT) tetramers (**a**) or K^b D227K (Mut) tetramers (**b**), and tetramer binding was quantified by flow cytometry. A representative experiment ($n \geq 6$) is shown. **c**, Pre-selection thymocytes were stained with tetramers as in **a**, and then incubated with monoclonal antibody to either V α 2 or CD8 β to prevent rebinding of dissociating tetramers. Tetramer binding is shown as the natural log (ln) of the geometric MFI versus time ($n \geq 3$).

Table 1 | Quantification of tetramer binding

Peptide	K_d WT (nM)*	K_d mutant (nM)*	$t_{1/2}$ WT (s)†	$t_{1/2}$ mutant (s)†	Selection
OVA	3.7 ± 0.7	39 ± 9.6	257 ± 35	80 ± 19	Negative
Q4	29 ± 7.2	241 ± 52	99 ± 39	31 ± 8	Negative
Q4R7	48 ± 9.5	282 ± 49	79 ± 19	20 ± 8	Negative
T4	55 ± 10.1	576 ± 97	73 ± 26	<12	Border
Q4H7	51 ± 9.1	721 ± 51	61 ± 16	<12	Positive
VSV	490 ± 89	No binding	ND	ND	None

* K_d values were determined from nonlinear regression analysis of tetramer binding curves with pre-selection double-positive thymocytes at 37 °C. The '±' value represents the 95% confidence interval from the nonlinear regression analysis.

† Half-life ($t_{1/2}$) values were calculated as described in Methods.

presence of antibodies that block rebinding to CD8 or the TCR (Fig. 2c). OVA and Q4R7 tetramers dissociated faster when their rebinding to the TCR was blocked. The dissociation of the T4 tetramer was nearly equal under both conditions, whereas tetramers of the positive selecting ligands Q4H7 and G4 dissociated more quickly when rebinding to CD8 was blocked. Taken together, these data indicate that binding of negative selecting ligands is dominated by the TCR, whereas binding of positive selecting ligands is more dependent on CD8; they also highlight the origin of the observed binding differences.

Signalling kinetics at the selection threshold

Given these subtle differences, we wanted to determine how positive and negative selecting ligands induce divergent signals. K^b tetramers were used to stimulate pre-selection OT-I double-positive thymocytes. Tetramer concentration was adjusted to achieve equal thymocyte occupancy for each ligand (Fig. 3a–e). OVA tetramers induced maximal Ca^{2+} flux, whereas the null ligand gave no response. The negative selector Q4R7 induced a rapid flux, whereas the positive selector Q4H7 induced Ca^{2+} flux at a slower rate (Fig. 3b). The 23-kDa phosphorylated species of CD3- ζ (p23- ζ) was efficiently induced by stimulation with OVA tetramer. Although the kinetics remained similar, induction of p23- ζ and ZAP-70 phosphorylation decreased incrementally in line with ligand potency (OVA > Q4 > Q4R7 > T4 > Q4H7; Fig. 3c, e, and data not shown). The induction of p23- ζ and ZAP-70 phosphorylation by VSV resembled the non-stimulated control.

Induction of LAT phosphorylation was quantitatively and kinetically distinct between negative and positive selectors (Fig. 3d, e). The negative selectors OVA, Q4 and Q4R7 induced a strong early peak of LAT phosphorylation that gradually declined over time. In contrast, T4 and the positive selector Q4H7 were weak inducers of LAT phosphorylation at all time points and the peak of LAT phosphorylation was delayed. Stimulation with OVA induced strong and transient ERK phosphorylation, peaking at ~1 min. The peak of ERK phosphorylation for the negative selectors Q4 and Q4R7 was slightly delayed, and the amount of phosphorylation induced correlated with the potency of the ligands (OVA > Q4 > Q4R7). However, the positive selector Q4H7 induced less phosphorylation of ERK more slowly. Thus, beginning with LAT, positive and negative selectors induced signals in a quantitatively and kinetically distinct manner. This observation cannot be explained by differences in occupancy because the thymocytes were loaded with equivalent amounts of ligand.

The spatial compartmentalization of Ras

These data raise the issue of how, at the selection border, a small increase in ligand affinity leads to subtle changes in p23- ζ and ZAP-70 phosphorylation, but to marked changes in downstream signalling events. Studies suggest that the kinetics of ERK induction is dependent on the subcellular localization of both ERK and its upstream intermediates^{35,36}. To examine the role of cellular localization of signalling molecules in thymic selection, pre-selection OT-I double-positive thymocytes were stimulated as in Fig. 3 and analysed by confocal microscopy. On stimulation, ZAP-70 is

recruited to the plasma membrane through its association with phosphorylated CD3- ζ .

We found that 95% of OVA-stimulated thymocytes induced recruitment of ZAP-70 phosphorylated at residue 319 to the plasma membrane within 2 min (full, 54%; partial, 41%). With Q4R7, 89% of cells showed plasma membrane recruitment of phosphorylated ZAP-70 (full, 56%; partial, 43%). In contrast, the positive selector Q4H7 induced much less localization of phosphorylated ZAP-70 to the plasma membrane (full, 12%; partial, 31%) and these differences remained after 10 min (Fig. 4a and Supplementary Fig. 2). Specific residues of phosphorylated LAT recruit Grb2–SOS, an adapter and guanine nucleotide exchange factor (GEF) module important for Ras activation in negative selection^{37,38}. The negative selectors OVA and Q4R7 efficiently recruited both Grb2 (100% OVA, 97% Q4R7) and SOS (99% OVA, 100% Q4R7) to the plasma membrane. None (0%) of the thymocytes stimulated with the positive selector Q4H7 showed this pattern of plasma membrane recruitment of Grb2–SOS (Fig. 4b).

Phospholipase C γ 1 is also recruited to phosphorylated LAT and, once activated, generates diacylglycerol (DAG)³⁷. DAG activates RasGRP1, a GEF important for Ras activation in positive selection³⁹. The negative selectors OVA and Q4R7 recruited RasGRP1, Ras and Raf-1 to the plasma membrane in 100% of the thymocytes analysed. In contrast, these molecules colocalized to the Golgi when stimulated with the positive selector Q4H7 (Fig. 4c, d, f, and Supplementary Figs 2 and 3). The trafficking patterns of Ras are dependent on acylation

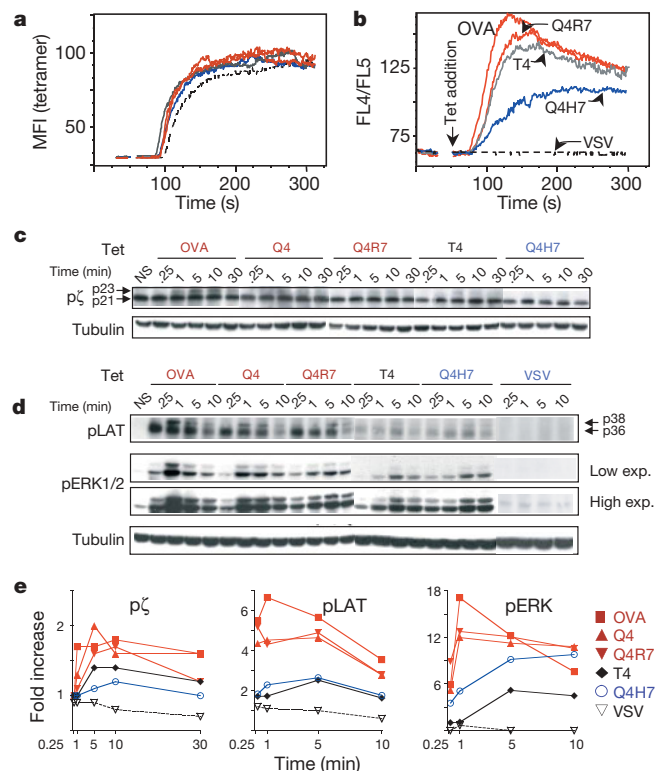


Figure 3 | Differences in TCR proximal signal induction distinguish positive and negative selecting ligands. **a**, Pre-selection OT-I double-positive thymocytes were incubated with concentrations of various peptide- K^b tetramers that gave equivalent occupancy. **b**, Ca^{2+} flux was measured by flow cytometry. **c**, **d**, Induction of phosphorylation of p23- ζ (**c**) LAT and ERK (**d**) was determined by western blot analysis of lysates of OT-I pre-selection double-positive thymocytes, stimulated with tetramers as in **a**. Two exposures are shown for the phosphorylated ERK blots (**d**). Antibody to tubulin was used as a loading control (**c**, **d**). **e**, Kinetics of the induction of p23- ζ , LAT and ERK phosphorylation is shown as the fold increase over nonstimulated cells using antibody to either tubulin (phosphorylated CD3- ζ and LAT) or ERK-2 (phosphorylated ERK) as a loading control.

because inhibition of the palmitoylation cycle with 2-bromopalmitate blocked its movement⁴⁰ (data not shown). These data also suggest that Ras is active in both compartments because only the activated form of Ras can recruit Raf-1 (ref. 36).

The negative selectors also targeted phosphorylated ERK to the plasma membrane within 2 min, where it remained concentrated at 10 min. In contrast, the positive selecting ligands induced an accumulation of phosphorylated ERK throughout the cell, especially at later times. With both types of selecting ligand, phosphorylated JNK accumulated throughout the cell with similar kinetics (Fig. 4e and data not shown)⁴¹. The net result is that only negative selectors induce segregation of phosphorylated ERK and phosphorylated JNK. Therefore, a key distinction between positive and negative selecting ligands is how they compartmentalize the MAPK signalling intermediates that have a role in regulating cell fate decisions.

To test the idea that signalling and compartmentalization patterns predict selection outcome, we studied the OVA K^b mutant ligand. We considered that the loss of CD8 participation observed with this ligand should decrease the potency of a negative selector below the selection threshold and convert it into a positive selector, as shown in CD8 α -null mice⁴². OVA presented by K^b mutant APCs induced CD69 expression with a potency in the range of positive selectors (Supplementary Fig. 4), and OVA K^b mutant tetramers induced Ca²⁺ flux similar to that induced by Q4H7 K^b wild-type tetramers. OVA K^b mutant and Q4H7 K^b wild-type tetramers induced LAT and ERK phosphorylation with similar kinetics (Fig. 5a–c). Furthermore, OVA K^b mutant tetramers induced localization patterns of Grb2, SOS, RasGRP1, Ras, Raf-1, phosphorylated ERK and phosphorylated JNK that closely resembled those induced by positive selectors (Fig. 4 and Supplementary Figs 2 and 3).

To confirm that this signalling profile leads to *bona fide* positive selection, we established reaggregated thymic organ cultures with pre-selection OT-I double-positive thymocytes, B6 Rag^{-/-} β 2m^{-/-} fetal thymic epithelium, and T2 wild-type K^b or T2 mutant-K^b APCs pulsed with Q4H7 or OVA as a source of MHC molecules⁴³. In

cultures lacking peptide, thymocytes remained double positive, whereas Q4H7-pulsed T2 wild-type K^b APCs induced positive selection (Fig. 5d). Reaggregates with OVA-pulsed T2 wild-type K^b APCs generated about 20-fold fewer cells, indicative of negative selection; however, cultures containing T2 mutant K^b APCs pulsed with OVA induced robust positive selection. These results argue that ligand potency establishes a compartmentalization pattern of the signalling machinery that predicts selection outcome.

Discussion

Our data provide insight into central tolerance by defining the boundary between positive and negative selection and by describing how thymocytes interpret ligands lying close to this boundary to generate differences in TCR proximal signals. Although pre-selection thymocytes respond to a wide range of ligands, the transition from positive to negative selection occurs across an extremely narrow range of ligand potency. The border peptide T4 functions as a negative selector at high concentrations, behaves as a positive selector at physiological doses, and is a poor antigen for peripheral OT-I T cells at all doses. In contrast, Q4R7, which is only 1.5 times more potent than T4, functions as a negative selector at all concentrations. Therefore, there is only a narrow window in which self-reactive thymocytes can potentially escape negative selection. These data document a sharp selection threshold that is established in the thymus, maintained in the periphery, and emphasizes the fidelity of central tolerance.

The binding characteristics of ligands comprising the selection boundary are strikingly similar. Only after analysing the individual components of thymocyte–tetramer interactions are the subtle differences between these ligands revealed. At the selection threshold, negative selecting ligands have a slightly higher TCR affinity than their positive selecting counterparts, consistent with previous studies^{12–15}. At the selection border, the TCR–pMHC affinity is similar to the CD8–pMHC affinity. In the context of kinetic proofreading, we can imagine that interactions longer than the lifetime of the

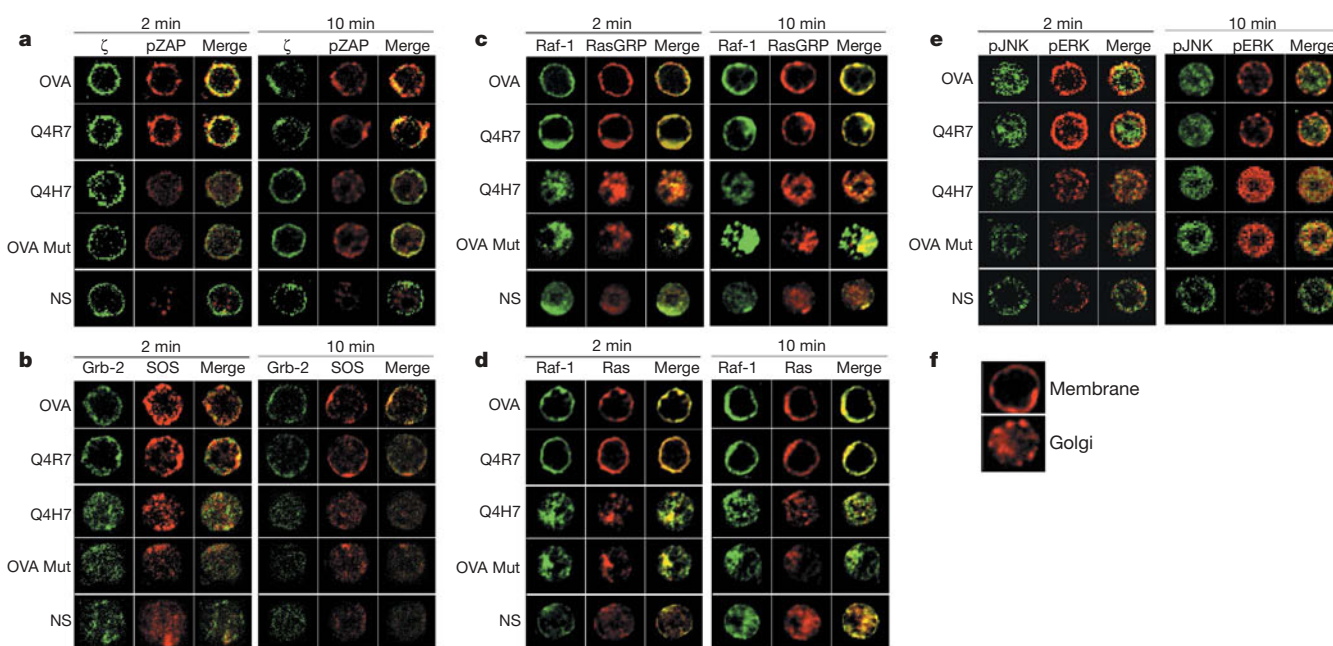


Figure 4 | Differential compartmentalization of early signal intermediates by positive and negative selecting ligands. Confocal analysis was done on pre-selection OT-I double-positive thymocytes stimulated, as in Fig. 3, for the indicated times and subjected to intracellular staining. **a**, CD3- ζ (green) and ZAP-70 phosphorylated on residue 319 (red); **b**, Grb2 (green) and SOS (red); **c**, Raf-1 (green) and RasGRP1 (red); **d**, Raf-1 (green) and Ras (red);

e, phosphorylated JNK (green) and phosphorylated ERK (red). **f**, Membrane-specific (LAT) and Golgi-specific (γ -1 adaptin) stains are shown as localization controls. The cells shown in each panel are representative of $n \geq 70$ tetramer-bound cells from three independent experiments. NS, non-stimulated.

CD8–pMHC complex allow Lck (carried by CD8) sufficient access to the TCR–CD3 complex to generate the signals necessary to induce negative selection, whereas shorter interactions lead to positive selection. Heterologous competition assays and the differential contributions of CD8 and TCR to ligand binding suggest, however, that negative selecting ligands may enhance the formation of the tripartite CD8–TCR–pMHC complex⁴⁴. Therefore, other parameters of ligand engagement and/or the geometrical orientation within the tripartite complex may also be involved in establishing differences in thymocyte signalling.

The mechanism used by thymocytes to convert small changes in analogue input (affinity) to a digital output (positive versus negative selection) begins very early in the TCR signalling cascade. Slight increases in affinity across the selection threshold generate incremental increases in the induction of p23- ζ and ZAP-70 phosphorylation;

however, negative selectors induce more efficient concentration of phosphorylated ZAP-70 at the membrane. This observation suggests that enhanced phosphorylation of LAT induced by negative selectors may be explained by an increase in the local concentration of this kinase. This leads to activation and translocation of RasGRP1 from the cytosol to the plasma membrane and recruitment of Grb2–SOS to phosphorylated LAT, resulting in activation of the Ras/Raf-1/ERK cascade at the plasma membrane. In stark contrast, positive selectors have no effect on the localization of Grb2–SOS, but induce recruitment and activation of RasGRP1, Ras, Raf-1 and ERK at the Golgi. Of note, plasma-membrane-localized Ras, Raf-1 and ERK are extremely sensitive to low signal input³⁵. Moreover, the activity of plasma-membrane-localized Ras is negatively regulated by the Ca^{2+} -dependent RasGAP, CAPRI^{36,45}, which may account for its transient activation. The kinetics of activation of the Ras/Raf-1/ERK cascade at the Golgi is slower and not subjected to stringent negative regulation^{35,36,45}. Taken together, these data suggest a mechanism underlying the kinetic differences in ERK activation observed with positive and negative selecting ligands^{22,41}. Considering the competing roles of phosphorylated ERK and phosphorylated JNK in thymic selection^{21,26}, the retention of ERK at the plasma membrane may enable phosphorylated JNK and other effector molecules to initiate negative selection successfully. These events are the first steps in a process that must continue for some time to achieve a final commitment to selection^{46,47}.

METHODS

Mice. All mice were bred in our colony and handled in accordance with Swiss laws.

CD69 assays. APCs expressing wild-type K^b or D227K mutant K^b MHC (a gift from T. Potter) were pulsed with peptide and incubated with pre-selection OT-I double-positive thymocytes (a ratio of one APC to two thymocytes) for 16 h and analysed by flow cytometry with Tree Star software. CD69 dose response curves were analysed by nonlinear regression using GraphPad software.

FTOC, Ca^{2+} flux and MHC I tetramers. FTOC assays, Ca^{2+} flux assays, and tetramer production were done as described^{29,32}.

Quantification of tetramer binding. Pre-selection OT-I double-positive thymocytes were stained with tetramers as described³³. K_d was determined by nonlinear regression analysis of the geometric mean fluorescence intensity (MFI) versus tetramer concentration curves (see Supplementary Information for details). Tetramer rebinding was blocked with antibody to K^b (Y3), V α 2 (B20) or CD8 β (53.5.8)³². Half-lives were calculated as described³³.

Western blotting. Thymocytes were stimulated with pMHC tetramers at 37 °C, lysates were subjected to western blot analysis, and bands were quantified by densitometry^{48,49}. An anti-phosphotyrosine monoclonal antibody (4G10) was used to detect phosphorylated CD3- ζ and phosphorylated LAT; polyclonal antibodies were used to detect phosphorylated ERK1/2. The identity of phosphorylated proteins was confirmed with antibody to LAT or ERK2, or rabbit antiserum to CD3- ζ (a gift from B. Alarcon).

Confocal microscopy. Tetramer stimulated double-positive thymocytes were stained, examined by confocal microscopy and quantified with LSM510 META software (Carl Zeiss) and ImageJ (NIH)⁴⁸. Blind analyses of localization patterns for each condition, performed by six individuals, determined cells representing the majority population ($n > 70$; see Supplementary Information for details).

Reaggregate thymic cultures. Reaggregate cultures were established from embryonic day (E15) thymic epithelial cells (TECs) from B6 $\text{Rag}^{-/-}\beta 2\text{m}^{-/-}$ embryos, OT-I pre-selection double-positive thymocytes and T2 APCs at a ratio of 10/30/1 (TECs/thymocytes/T2 APCs) and analysed as described⁵⁰.

Received 6 July; accepted 18 September 2006.

Published online 1 November 2006.

1. Starr, T. K., Jameson, S. C. & Hogquist, K. A. Positive and negative selection T cells. *Annu. Rev. Immunol.* 21, 139–176 (2003).
2. Germain, R. N. The T cell receptor for antigen: signaling and ligand discrimination. *J. Biol. Chem.* 276, 35223–35226 (2001).
3. Sebzda, E. *et al.* Positive and negative thymocyte selection induced by different concentrations of a single peptide. *Science* 263, 1615–1618 (1994).
4. Ashton-Rickardt, P. G. *et al.* Evidence for a differential avidity model of T cell selection in the thymus. *Cell* 76, 651–663 (1994).
5. Fontenot, J. D. & Rudensky, A. Y. A well adapted regulatory contrivance: regulatory T cell development and the Forkhead family transcription factor Foxp3. *Nature Immunol.* 6, 331–337 (2005).

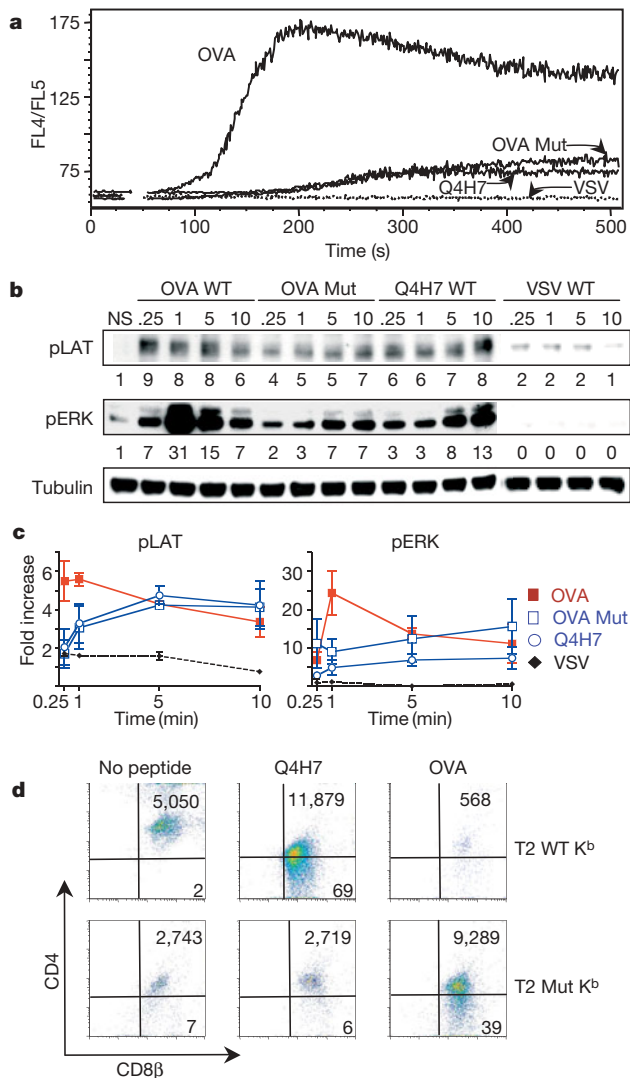


Figure 5 | Signalling and compartmentalization profiles are predictive of selection outcome. **a**, OT-I pre-selection double-positive thymocytes were stimulated with tetramers and the Ca^{2+} flux (**a**) and pattern of LAT and ERK phosphorylation (**b**) was determined as in Fig. 3. Numbers below the blots indicate the fold induction. **c**, Induction kinetics for phosphorylated LAT and phosphorylated ERK, determined as in Fig. 3 (mean $\pm$ s.d.; $n = 4$). **d**, Reaggregate thymic organ cultures containing E15.5 TECs from $\text{Rag}^{-/-}\beta 2\text{m}^{-/-}$ mice, OT-I pre-selection double-positive thymocytes and either T2 wild-type K^b or T2 D227K (Mut) K^b peptide-pulsed APCs. The number of thymocytes recovered in each reaggregate culture (top right) and the percentage of CD8 $\alpha\beta$ single-positive thymocytes (bottom right) are indicated.

6. Sakaguchi, S. Naturally arising Foxp3-expressing CD25⁺CD4⁺ regulatory T cells in immunological tolerance to self and non-self. *Nature Immunol.* **6**, 345–352 (2005).
7. Leishman, A. J. *et al.* Precursors of functional MHC class I- or class II-restricted CD8 $\alpha\alpha$ ⁺ T cells are positively selected in the thymus by agonist self-peptides. *Immunity* **16**, 355–364 (2002).
8. Yamagata, T., Mathis, D. & Benoist, C. Self-reactivity in thymic double-positive cells commits cells to a CD8 $\alpha\alpha$ lineage with characteristics of innate immune cells. *Nature Immunol.* **5**, 597–605 (2004).
9. Hogquist, K. A., Baldwin, T. A. & Jameson, S. C. Central tolerance: learning self-control in the thymus. *Nature Rev. Immunol.* **5**, 772–782 (2005).
10. McKeithan, T. W. Kinetic proofreading in T-cell receptor signal transduction. *Proc. Natl Acad. Sci. USA* **92**, 5042–5046 (1995).
11. Goldstein, B., Faeder, J. R. & Hlavacek, W. S. Mathematical and computational models of immune-receptor signalling. *Nature Rev. Immunol.* **4**, 445–456 (2004).
12. Alam, S. M. *et al.* T-cell receptor affinity and thymocyte positive selection. *Nature* **381**, 616–620 (1996).
13. Davis, M. M. *et al.* Ligand recognition by $\alpha\beta$ T cell receptors. *Annu. Rev. Immunol.* **16**, 523–544 (1998).
14. Liu, C. P., Crawford, F., Marrack, P. & Kappler, J. T cell positive selection by a high density, low affinity ligand. *Proc. Natl Acad. Sci. USA* **95**, 4522–4526 (1998).
15. Williams, C. B., Engle, D. L., Kersh, G. J., Michael White, J. & Allen, P. M. A kinetic threshold between negative and positive selection based on the longevity of the T cell receptor–ligand complex. *J. Exp. Med.* **189**, 1531–1544 (1999).
16. Garcia, K. C. *et al.* CD8 enhances formation of stable T-cell receptor/MHC class I molecule complexes. *Nature* **384**, 577–581 (1996).
17. Wyer, J. R. *et al.* T cell receptor and coreceptor CD8 $\alpha\alpha$ bind peptide–MHC independently and with distinct kinetics. *Immunity* **10**, 219–225 (1999).
18. Zamoyska, R. *et al.* The influence of the Src-family kinases, Lck and Fyn, on T cell differentiation, survival and activation. *Immunol. Rev.* **191**, 107–118 (2003).
19. Daniels, M. A. *et al.* CD8 binding to MHC class I molecules is influenced by T cell maturation and glycosylation. *Immunity* **15**, 1051–1061 (2001).
20. van Leeuwen, J. E. & Samelson, L. E. T cell antigen-receptor signal transduction. *Curr. Opin. Immunol.* **11**, 242–248 (1999).
21. Alberola-Ila, J. & Hernandez-Hoyos, G. The Ras/MAPK cascade and the control of positive selection. *Immunol. Rev.* **191**, 79–96 (2003).
22. Werlen, G., Hausmann, B., Naehrer, D. & Palmer, E. Signaling life and death in the thymus: timing is everything. *Science* **299**, 1859–1863 (2003).
23. Palmer, E. Negative selection—clearing out the bad apples from the T-cell repertoire. *Nature Rev. Immunol.* **3**, 383–391 (2003).
24. Bommhardt, U., Scheuring, Y., Bickel, C., Zamoyska, R. & Hunig, T. MEK activity regulates negative selection of immature CD4⁺CD8⁺ thymocytes. *J. Immunol.* **164**, 2326–2337 (2000).
25. Mariathasan, S. *et al.* Duration and strength of extracellular signal-regulated kinase signals are altered during positive versus negative thymocyte selection. *J. Immunol.* **167**, 4966–4973 (2001).
26. Dong, C., Davis, R. J. & Flavell, R. A. MAP kinases in the immune response. *Annu. Rev. Immunol.* **20**, 55–72 (2002).
27. Alam, S. M. *et al.* Qualitative and quantitative differences in T cell receptor binding of agonist and antagonist ligands. *Immunity* **10**, 227–237 (1999).
28. Hare, K. J., Jenkinson, E. J. & Anderson, G. CD69 expression discriminates MHC-dependent and -independent stages of thymocyte positive selection. *J. Immunol.* **162**, 3978–3983 (1999).
29. Hogquist, K. A. *et al.* T cell receptor antagonist peptides induce positive selection. *Cell* **76**, 17–27 (1994).
30. Savage, P. A., Boniface, J. J. & Davis, M. M. A kinetic basis for T cell receptor repertoire selection during an immune response. *Immunity* **10**, 485–492 (1999).
31. Crawford, F., Kozono, H., White, J., Marrack, P. & Kappler, J. Detection of antigen-specific T cells with multivalent soluble class II MHC covalent peptide complexes. *Immunity* **8**, 675–682 (1998).
32. Daniels, M. A. & Jameson, S. C. Critical role for CD8 in T cell receptor binding and activation by peptide/major histocompatibility complex multimers. *J. Exp. Med.* **191**, 335–346 (2000).
33. Rosette, C. *et al.* The impact of duration versus extent of TCR occupancy on T cell activation: a revision of the kinetic proofreading model. *Immunity* **15**, 59–70 (2001).
34. Potter, T. A., Rajan, T. V., Dick, R. F. II & Bluestone, J. A. Substitution at residue 227 of H-2 class I molecules abrogates recognition by CD8-dependent, but not CD8-independent, cytotoxic T lymphocytes. *Nature* **337**, 73–75 (1989).
35. Harding, A., Tian, T., Westbury, E., Frische, E. & Hancock, J. F. Subcellular localization determines MAP kinase signal output. *Curr. Biol.* **15**, 869–873 (2005).
36. Mor, A. & Philips, M. R. Compartmentalized Ras/MAPK signaling. *Annu. Rev. Immunol.* **24**, 771–800 (2006).
37. Sommers, C. L., Samelson, L. E. & Love, P. E. LAT: a T lymphocyte adapter protein that couples the antigen receptor to downstream signaling pathways. *BioEssays* **26**, 61–67 (2004).
38. Gong, Q. *et al.* Disruption of T cell signaling networks and development by Grb2 haploid insufficiency. *Nature Immunol.* **2**, 29–36 (2001).
39. Dower, N. A. *et al.* RasGRP is essential for mouse thymocyte differentiation and TCR signaling. *Nature Immunol.* **1**, 317–321 (2000).
40. Rocks, O. *et al.* An acylation cycle regulates localization and activity of palmitoylated Ras isoforms. *Science* **307**, 1746–1752 (2005).
41. Werlen, G., Hausmann, B. & Palmer, E. A motif in the $\alpha\beta$ T-cell receptor controls positive selection by modulating ERK activity. *Nature* **406**, 422–426 (2000).
42. Goldrath, A. W., Hogquist, K. A. & Bevan, M. J. CD8 lineage commitment in the absence of CD8. *Immunity* **6**, 633–642 (1997).
43. Hare, K. J., Jenkinson, E. J. & Anderson, G. *In vitro* models of T cell development. *Semin. Immunol.* **11**, 3–12 (1999).
44. Yachi, P. P., Ampudia, J., Gascoigne, N. R. & Zal, T. Nonstimulatory peptides contribute to antigen-induced CD8–T cell receptor interaction at the immunological synapse. *Nature Immunol.* **6**, 785–792 (2005).
45. Bivona, T. G. *et al.* Phospholipase C γ activates Ras on the Golgi apparatus by means of RasGRP1. *Nature* **424**, 694–698 (2003).
46. Liu, X. *et al.* Restricting Zap70 expression to CD4⁺CD8⁺ thymocytes reveals a T cell receptor-dependent proofreading mechanism controlling the completion of positive selection. *J. Exp. Med.* **197**, 363–373 (2003).
47. McNeil, L. K., Starr, T. K. & Hogquist, K. A. A requirement for sustained ERK signaling during thymocyte positive selection *in vivo*. *Proc. Natl Acad. Sci. USA* **102**, 13574–13579 (2005).
48. Teixeira, E. *et al.* T cell division and death are segregated by mutation of TCR β chain constant domains. *Immunity* **21**, 515–526 (2004).
49. Teixeira, E., Fuentes, P., Galocha, B., Alarcon, B. & Bragado, R. T cell receptor-mediated signal transduction controlled by the β chain transmembrane domain: apoptosis-deficient cells display unbalanced mitogen-activated protein kinases activities upon T cell receptor engagement. *J. Biol. Chem.* **277**, 3993–4002 (2002).
50. Anderson, G., Jenkinson, E. J., Moore, N. C. & Owen, J. J. MHC class II-positive epithelium and mesenchyme cells are both required for T-cell development in the thymus. *Nature* **362**, 70–73 (1993).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank R. Clark and T. Potter for hospitality and for providing OT-I Rag^{-/-} β 2m^{-/-} mice; S. Jameson for K^b plasmids; G. de Libero and A. Schrum for reading the manuscript; V. Jägglin for assistance with the Ca²⁺ flux analysis; and E. Wagner and W. Hänggi for animal husbandry. This work was supported by grants from Novartis, the Swiss National Science Foundation, the US Cancer Research Institute (to M.A.D. and K.H.), Hoffmann La Roche, and the NIH (to N.R.J.G.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.P. (ed.palmer@unibas.ch).

LETTERS

Active galactic nuclei as scaled-up Galactic black holes

I. M. McHardy¹, E. Koering¹, C. Knigge¹, P. Uttley² & R. P. Fender¹

A long-standing question is whether active galactic nuclei (AGN) vary like Galactic black hole systems when appropriately scaled up by mass^{1–3}. If so, we can then determine how AGN should behave on cosmological timescales by studying the brighter and much faster varying Galactic systems. As X-ray emission is produced very close to the black holes, it provides one of the best diagnostics of their behaviour. A characteristic timescale—which potentially could tell us about the mass of the black hole—is found in the X-ray variations from both AGN and Galactic black holes^{1–6}, but whether it is physically meaningful to compare the two has been questioned⁷. Here we report that, after correcting for variations in the accretion rate, the timescales can be physically linked, revealing that the accretion process is exactly the same for small and large black holes. Strong support for this linkage comes, perhaps surprisingly, from the permitted optical emission lines in AGN whose widths (in both broad-line AGN and narrow-emission-line Seyfert 1 galaxies) correlate strongly with the characteristic X-ray timescale, exactly as expected from the AGN black hole masses and accretion rates. So AGN really are just scaled-up Galactic black holes.

The first detailed observations of AGN X-ray variability showed scale-invariant behaviour on all timescales from approximately days to minutes^{8,9}, with no characteristic timescale from which black hole masses (M_{BH}) might be deduced. However, subsequent observations^{1–3} showed that, on longer timescales, a characteristic timescale could be derived from the power spectral densities (PSDs; that is, variability power, $P(\nu)$, at frequency, ν , or timescale, $1/\nu$) of the X-ray light curves.

All AGN PSDs are best fitted on long timescales by a power-law of slope -1 ($P(\nu) \propto \nu^{-\alpha}$ with $\alpha \approx 1$) which breaks to a steeper slope ($\alpha > 2$) on timescales shorter than a 'break' timescale, T_{B} . For some AGN, the $\alpha \approx 1$ slope can be followed to long timescales for >3 decades with no further break, similar to Galactic black hole X-ray binary systems (GBHs) in their 'soft' states^{5,7,10–13}. For other AGN, the slope can only be followed for <2 decades, which is insufficient to distinguish them from GBHs in their 'hard' states where, in the power-law description of the PSD, a second break, to slope $\alpha \approx 0$, is seen ~ 1.5 – 2 decades below the $\alpha \approx 1$ – 2 break. Here we use the timescale associated with the $\alpha \approx 1$ – 2 break as a characteristic timescale, irrespective of likely state. The reason for the sudden decrease in variability power on timescales shorter than T_{B} is not clear, but the variability probably originates within the accretion disk¹⁴ surrounding the black hole and T_{B} may be associated with the inner edge of the disk.

A major difficulty in establishing a quantitative timing link between AGN and GBHs has been the large scatter in the M_{BH} – T_{B} relationship⁷. In particular, for a given M_{BH} , the high accretion rate narrow line Seyfert 1 galaxies (NLS1s) have smaller values of T_{B} than other AGN^{5,11}. We therefore suggested that T_{B} is inversely dependent on a second variable, possibly the black hole spin, but probably the accretion rate^{5,11,12} (often written $\dot{m}_{\text{E}}$, in units of the maximum possible,

Eddington, accretion rate). Other researchers, estimating T_{B} from the excess variance in the X-ray light curves, also find a correlation of T_{B} with M_{BH} but the dependence on $\dot{m}_{\text{E}}$ is not clear^{6,13}. Here, we properly quantify the relationship between T_{B} , M_{BH} and $\dot{m}_{\text{E}}$. To make the parameters properly independent, we fit to the observable quantity, the bolometric luminosity, L_{bol} , rather than $\dot{m}_{\text{E}}$, although for the objects discussed here, which are radio-quiet, $\dot{m}_{\text{E}}$ is well approximated by $L_{\text{bol}}/L_{\text{E}}$ (Eddington luminosity $L_{\text{E}} \propto M_{\text{BH}}$).

Motivated by the rough linear scaling between T_{B} and M_{BH} , and the rough decrease of T_{B} with increasing $\dot{m}_{\text{E}}$, we hypothesize that $\log T_{\text{B}} = A \log M_{\text{BH}} - B \log L_{\text{bol}} + C$ and determine the best-fit values of A , B and C from a simple parameter grid search. Details of PSD parameterization, fitting procedure and data relevant to this paper are given in the Supplementary Information. We first fit to the 10 AGN that have well measured PSD breaks and reasonable measurements of mass and bolometric luminosity¹². This sample contains a range of accretion rates from $\dot{m}_{\text{E}} > \sim 1$ (Akn 564) to $\sim 10^{-3}$ (NGC 4395). The fit (Fig. 1) is good with $A = 2.17^{+0.32}_{-0.25}$, $B = 0.90^{+0.3}_{-0.2}$ and $C = -2.42^{+0.22}_{-0.25}$.

To determine whether the same scaling relationship extends to GBHs, we include two bright GBHs (Cyg X-1 and GRS 1915+105) in radio-quiet states where—for proper comparison—their high frequency PSDs are well described by the same cut-off ('breaking') power-law model which best describes AGN^{10,15}, and where broad band X-ray flux provides a good measurement of bolometric luminosity. For Cyg X-1, we combined measurements of T_{B} (ref. 10) with simultaneous measurement of the bolometric luminosity¹⁶ over a range of luminosities. For GRS 1915+105, we measured an average T_{B} from the original X-ray data, and determined L_{bol} from the published fluxes¹⁵ and generally accepted distance (11 kpc).

Using Cyg X-1, we can test whether the AGN scaling relationship applies within any one object (that is, at fixed mass). We bin into 5 luminosity bins and find that $T_{\text{B}} \propto L_{\text{bol}}^{-1.3 \pm 0.2}$. This dependence is consistent with that for AGN, although slightly larger, probably because the bolometric range covered includes the transition zone between hard and soft states where more rapid changes of timescale with luminosity occur than in either state alone¹⁷.

It is particularly important to determine whether these Galactic data can be fitted, self-consistently, with the AGN data. We first fit to Cyg X-1 and the AGN and find excellent agreement with the AGN alone. Including also GRS 1915+105, whose break timescale is $\sim 10\times$ shorter than for Cyg X-1, we again find good agreement (Fig. 1) with $A = 2.10 \pm 0.15$, $B = 0.98 \pm 0.15$ and $C = -2.32 \pm 0.2$. As the fit is good, no unknown source of error need be invoked, implying that no other parameter (such as spin) has as large an effect on T_{B} as M_{BH} or $\dot{m}_{\text{E}}$. Thus we answer a long-standing question and show—using a self-consistently derived set of AGN and GBH timing data—that, over a range of $\sim 10^8$ in mass and $\sim 10^3$ in accretion rate, AGN behave just like scaled-up GBHs. Assuming $\dot{m}_{\text{E}} \approx L_{\text{bol}}/L_{\text{E}}$, then $T_{\text{B}} \approx M_{\text{BH}}^{1.12}/\dot{m}_{\text{E}}^{0.98}$.

¹School of Physics and Astronomy, The University, Southampton SO17 1BJ, UK. ²Astronomical Institute Anton Pannekoek, University of Amsterdam, Kruislaan 403, 1098 SJ Amsterdam, The Netherlands.

If the break timescale is proportional to a thermal or viscous timescale associated with the inner radius of the accretion disk, R_{disk} , then from our fit we expect $R_{\text{disk}} \propto \dot{m}_E^{-2/3}$. Models based on evaporation of the inner disk¹⁸ predict $R_{\text{disk}} \propto \dot{m}_E^{-0.85}$ and $T_B \propto M_{\text{BH}}^{1.2}$, which is not too far from our observations.

In their hard states, GBH PSDs are often better described by the combination of two or more lorentzian-shaped components¹⁹ whose timescales approximate the break timescales in the power-law description of the PSD. Although not included in our fits as a number of assumptions are required, it has already been shown, from combined radio and X-ray observations²⁰, that in the hard state of GBH GX339–4 the timescale associated with the lorentzian closest to the $\alpha \approx 1$ –2 break timescale varies as $\sim 1/\dot{m}_E$ (see the Supplementary Information for more details). (Combined radio and X-ray luminosities also underpin another, non-timing, analysis, from which a strong scaling, particularly of jet properties, between AGN and GBHs has recently been shown^{21,22}.)

In Fig. 2 we show a projection of the T_B – M_{BH} – $\dot{m}_E$ plane and all objects lie close to it. Thus the same process (for example, accretion rate variations¹⁴) probably produces X-ray variations in the same way in all accreting black holes. Thus, although we agree completely with previous criticism⁷ that T_B , on its own, is not a precise indicator of M_{BH} , we show that T_B , when combined with an estimate of $\dot{m}_E$, is a very good mass indicator. This indicator may have particular value where mass determination is otherwise difficult—for obscured AGN and for potential intermediate mass (10^3 – 10^4 solar masses) black holes, for example.

Our results establish that T_B is a powerful tracer of the innermost accretion processes in AGN and GBHs. These processes produce photons that must affect the larger scale AGN properties, but no

relationship has yet been found between T_B and such properties. One particularly important property, often used to classify AGN, is the width of the permitted optical emission lines. These lines are narrower in more X-ray variable AGN^{23,24}, but the reason has never been satisfactorily explained.

Theoretically, a correlation between M_{BH} , $\dot{m}_E$ and linewidth is expected from simple scaling relationships for the gas surrounding the black hole (the broad line region, BLR) from which emission lines, broadened by Doppler velocities, V , originate. We assume that the emission lines result from photoionization and that the ionizing luminosity $L \propto M_{\text{BH}}\dot{m}_E$. We also assume virial motion for the BLR gas: that is, $GM_{\text{BH}}/R_{\text{BLR}} \approx V^2$, where R_{BLR} is the inner radius of the BLR, and $R_{\text{BLR}} \propto L^a$. For the ‘locally optimized condition’ for production of emission lines by gas at the same optimum density and ionization state, we expect $a = 0.5$, and the most recent observational study²⁵ finds $a = 0.518 \pm 0.039$. If $a = 0.5$, we expect $V^4 \approx M_{\text{BH}}/\dot{m}_E$.

A strong independent test of our derivation that $T_B \propto M_{\text{BH}}/\dot{m}_E$ is therefore provided by plotting (Fig. 3) T_B versus V , for all 9 of our primary sample of 10 AGN with well measured PSD breaks, for which measurements of the FWHM of the variable, broad component, of the H β line, V , exist²⁶. We note a much tighter relationship than in any previous correlations between linewidth and other parameters quantifying variability—for example, fractional variability of the optical continuum²⁷ or r.m.s. X-ray variability^{23,24}. Both T_B and V are observables, so this relationship depends on no assumptions.

Parameterizing $\log T_B = D \log V + E$, we perform a grid search for the best-fit values of D and E . Measurement errors on V are typically ~ 10 – 15% , but differences of $>50\%$ can occur between different

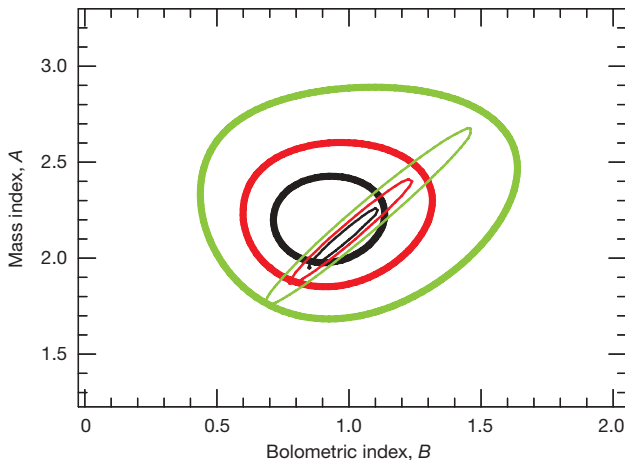


Figure 1 | Confidence contours for the mass and bolometric indices A and B . Here we show the 68% (black), 90% (red) and 95% (green) confidence contours for the dependence of the PSD break timescale on black hole mass and bolometric luminosity. The mass and bolometric indices, A and B , are defined in the main text. The thick contours refer to the primary sample of 10 AGN with well measured PSD breaks, that is NGC 3227, NGC 3516, NGC 3783, NGC 4051, NGC 4151, NGC 4395, NGC 5506, MCG-6-30-15, Mkn 766 and Akn 564. The best fit values for this fit are $A = 2.17^{+0.3}_{-0.2}$, $B = 0.90^{+0.25}_{-0.2}$ and $C = -2.42^{+0.22}_{-0.25}$ (1σ errors), reduced $\chi^2 = 0.85$, 7 degrees of freedom (d.o.f.), for T_B in days, M_{BH} in 10^6 solar masses and L_{bol} in 10^{44} erg s $^{-1}$. For a combined fit to the 10 AGN and the much lower mass Galactic black hole system Cyg X-1, we find $A = 2.10^{+0.23}_{-0.18}$, $B = 0.98^{+0.20}_{-0.16}$ and $C = -2.33 \pm 0.15$ (reduced $\chi^2 = 0.82$, 10 d.o.f.), which is excellent agreement with the AGN on their own. For Cyg X-1 we assume $M_{\text{BH}} = 15 \pm 5$ solar masses, but assuming $M_{\text{BH}} = 10$, or 20, changes each fit parameter by only 0.025. Including also the Galactic black hole system GRS 1915+105, we find $A = 2.10 \pm 0.15$, $B = 0.98 \pm 0.15$ and $C = -2.32 \pm 0.2$ (reduced $\chi^2 = 0.85$, 11 d.o.f.). The contours for this latter fit are shown as thin lines. Even at the 1σ (that is, 68% confidence) level, the contours for the AGN on their own, and for the combined AGN and GBH sample, completely overlap (as do the offset constants, C).

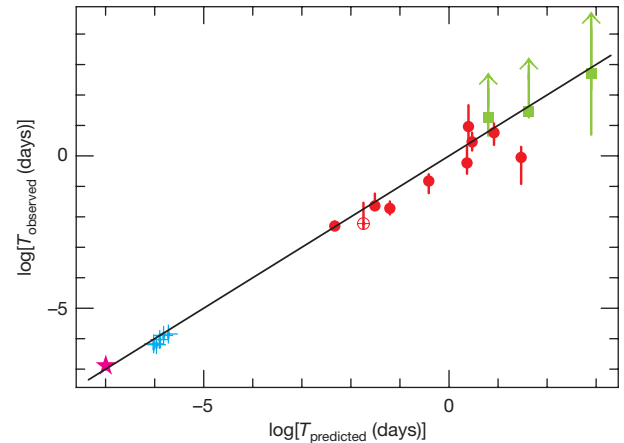


Figure 2 | Edge-on projection of our sample and the T_B – M_{BH} – L_{bol} plane. Here we show that the predicted break timescales, $T_{\text{predicted}}$, derived by inserting the observed bolometric luminosities and masses into the best fit relationship to the combined sample of AGN and GBHs shown in Fig. 1 (that is, $\log T_B = 2.1 \log M_{\text{BH}} - 0.98 \log L_{\text{bol}} - 2.32$), agree very well with the observed break timescales, T_{observed} , for all objects. Here we show GRS 1915+105 as a filled maroon star, Cyg X-1 as blue crosses and the 10 AGN as red circles. The low luminosity AGN (LLAGN), NGC 4395, is shown as an open crossed red circle; the other 9 AGN are filled red circles. Although not included in the fit as the upper limit on their break timescales are unbounded (hence the up-pointing green arrows), we also plot (filled green squares) NGC 5548, and Fairall 9 and the LLAGN NGC 4258. If the predicted and observed break timescales are identical, then an object will lie exactly on the black line, which is a projection of the best-fit three-dimensional T_B – M_{BH} – L_{bol} plane and is not a fit simply in two dimensions to the points shown in this figure. We plot the 90% confidence errors on T_{observed} which are usually quoted. In some cases these errors are too small to show beyond the symbols. The only noticeable outlier is NGC 5506 which, on the basis of a narrow Pa β near-infrared line, is classed as an obscured NLS1³⁰. However, its current mass estimate, less well determined than for most other AGN, implies $L_{\text{bol}}/L_E \approx 2.6\%$, which is surprisingly low for an NLS1. If the mass were a factor of ~ 5 lower, L_{bol}/L_E would be more normal for an NLS1 and NGC 5506 would sit directly on the plane.

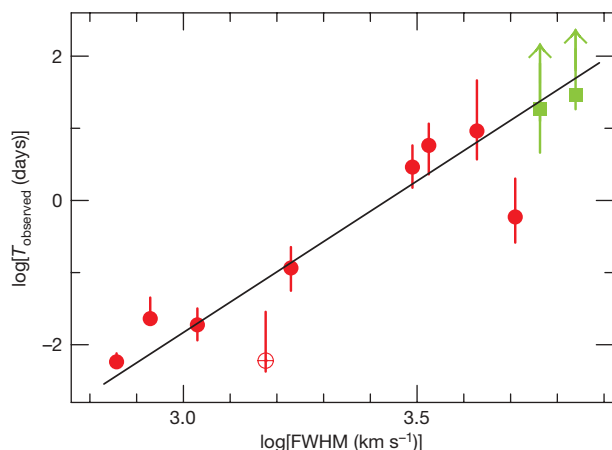


Figure 3 | Correlation of optical emission linewidth with PSD break timescale. We show that the full-width at half-maximum intensity (FWHM) of the H β optical emission line in AGN is strongly correlated with the observed PSD break timescale, T_{observed} . For $\log T_B = D \log V + E$, we find $D = 4.20^{+0.71}_{-0.56}$, $E = -14.43$ (reduced $\chi^2 = 1.20$, 7 d.o.f.), with T_B in days and width, V , in km s^{-1} . The fit is to the 9 AGN with both well measured PSD break timescales and measured values of V . As in Fig. 2, the LLAGN NGC 4395 is shown as an open crossed red circle and the other 8 AGN are filled red circles. NGC 5506 is present in Fig. 2 but missing here, as it is heavily obscured so H β is undetectable. As different lines may be produced at different distances from the black hole, alternative lines may not be used. Although not included in the fit, we plot (filled green squares, as in Fig. 2) Fairall 9 and NGC 5548 for whom linewidths are available but whose upper break timescales are unbounded. We note that their present values of T_B are consistent with the fit for the other AGN. Where available we use V as detected in the r.m.s. spectrum²⁶, as it is a better estimate of the variable component from the BLR and is less contaminated by constant narrow components originating further away. The H β broad linewidth of the LLAGN NGC 4395 is poorly determined. If it is removed from the sample, the fit improves slightly ($D = 3.98^{+0.64}_{-0.49}$, $E = -13.62$ with reduced $\chi^2 = 0.96$, 6 d.o.f.). The AGN with $\log \text{FWHM} = 3.72$ which lies below the best fit line is NGC 3227. For both NGC 3227 and NGC 4395, there are large time differences between the measurement of T_B and V which may introduce additional scatter. Error bars on T_B are 90% confidence, as in Fig. 2.

epochs²⁶. Thus large time differences between measurements of T_B and V can introduce additional scatter. Here we include a 30% error in linewidth, which is an estimate of the typical combined statistical and systematic uncertainty. Fitting to the primary sample of 9 AGN we find $D = 4.20^{+0.71}_{-0.56}$, strongly supporting our earlier derivation that $T_B \approx M_{\text{BH}}/\dot{m}_E$. Adopting slightly different values of the linewidth from other observers changes the fit very little. Thus two apparently quite different observable characteristics of AGN, that is, X-ray variability and optical linewidth, can both be explained as simply depending on $M_{\text{BH}}/\dot{m}_E$, thereby linking small scale nuclear accretion properties to larger scale AGN properties.

Our results, commented on in more detail in the Supplementary Information, have important implications for understanding the different types of active galaxy as differentiated by their optical linewidths, in particular the NLS1s^{28,29}. It is not M_{BH} or $\dot{m}_E$ on their own which define the linewidth, but the ratio $M_{\text{BH}}/\dot{m}_E$. The observed small masses of NLS1s are a selection effect as, for an AGN with M_{BH} greater than a few $\times 10^6$ solar masses to produce narrow emission lines, $\dot{m}_E$ must exceed the Eddington limit. Particular orientations, such as face-on, although not ruled out, are not required, nor is an unusual distribution or density of the surrounding gas. Apart from a lower ratio of $M_{\text{BH}}/\dot{m}_E$, NLS1s are no different to other AGN.

Received 24 May; accepted 24 October 2006.

1. McHardy, I. M. EXOSAT observations of variability in active galactic nuclei. *Mem. Soc. Astron. Ital.* **59**, 239–259 (1988).
2. Edelson, R. & Nandra, K. A cutoff in the X-ray fluctuation power density spectrum of the Seyfert 1 galaxy NGC 3516. *Astrophys. J.* **514**, 682–690 (1999).

3. Uttley, P., McHardy, I. M. & Papadakis, I. E. Measuring the broad-band power spectra of active galactic nuclei with RXTE. *Mon. Not. R. Astron. Soc.* **332**, 231–250 (2002).
4. Markowitz, A. *et al.* X-ray fluctuation power spectral densities of Seyfert 1 galaxies. *Astrophys. J.* **593**, 96–114 (2003).
5. McHardy, I. M., Papadakis, I. E., Uttley, P., Page, M. J. & Mason, K. O. Combined long and short time-scale X-ray variability of NGC 4051 with RXTE and XMM-Newton. *Mon. Not. R. Astron. Soc.* **348**, 783–801 (2004).
6. Papadakis, I. E. The scaling of the X-ray variability with black hole mass in active galactic nuclei. *Mon. Not. R. Astron. Soc.* **348**, 207–213 (2004).
7. Done, C. & Gierlinski, M. Scaling variability from stellar to supermassive black holes. *Mon. Not. R. Astron. Soc.* **364**, 208–216 (2005).
8. Lawrence, A., Watson, M. G., Pounds, K. A. & Elvis, M. Low-frequency divergent X-ray variability in the Seyfert galaxy NGC4051. *Nature* **325**, 694–696 (1987).
9. McHardy, I. M. & Czerny, B. Fractal X-ray time variability and spectral invariance of the Seyfert galaxy NGC5506. *Nature* **325**, 696–698 (1987).
10. Axelsson, M., Borgonovo, L. & Larsson, S. Evolution of the 0.01–25 Hz power spectral components in Cygnus X-1. *Astron. Astrophys.* **438**, 999–1012 (2005).
11. McHardy, I. M., Gunn, K. F., Uttley, P. & Goad, M. R. MCG-6-30-15: long time-scale X-ray variability, black hole mass and active galactic nuclei high states. *Mon. Not. R. Astron. Soc.* **359**, 1469–1480 (2005).
12. Uttley, P. & McHardy, I. M. X-ray variability of NGC 3227 and 5506 and the nature of active galactic nucleus ‘states’. *Mon. Not. R. Astron. Soc.* **363**, 586–596 (2005).
13. McClintock, J. E. & Remillard, R. A. in *Compact Stellar X-ray Sources* (eds Lewin, W. H. G. & van der Klis, M.) 157–213 (Cambridge Univ. Press, Cambridge, UK, 2006).
14. Lyubarskii, I. E. Flicker noise in accretion discs. *Mon. Not. R. Astron. Soc.* **292**, 679–685 (1997).
15. Trudolyubov, S. P. On the two types of steady hard X-ray states of GRS 1915+105. *Astrophys. J.* **558**, 276–282 (2001).
16. Wilms, J., Nowak, M. A., Pottschmidt, K., Pooley, G. & Fritz, S. Long term variability of Cygnus X-1. IV. Spectral evolution 1999–2004. *Astron. Astrophys.* **447**, 245–261 (2006).
17. Pottschmidt, K. *et al.* Long term variability of Cygnus X-1. I. X-ray spectral-temporal correlations in the hard state. *Astron. Astrophys.* **407**, 1039–1058 (2003).
18. Liu, B. F., Yuan, W., Meyer, F., Meyer-Hofmeister, E. & Xie, G. Z. Evaporation of accretion disks around black holes: The disk-corona transition and the connection to the advection-dominated accretion flow. *Astrophys. J.* **527**, L17–L20 (1999).
19. Nowak, M. A. Are there three peaks in the power spectra of GX 339-4 and Cyg X-1? *Mon. Not. R. Astron. Soc.* **318**, 361–367 (2000).
20. Migliari, S., Fender, R. P. & van der Klis, M. Correlation between radio luminosity and X-ray timing frequencies in neutron star and black hole X-ray binaries. *Mon. Not. R. Astron. Soc.* **363**, 112–120 (2005).
21. Merloni, A., Heinz, S. & di Matteo, T. A fundamental plane of black hole activity. *Mon. Not. R. Astron. Soc.* **345**, 1057–1076 (2003).
22. Falcke, H., Koerding, E. & Markoff, S. A scheme to unify low-power accreting black holes. Jet-dominated accretion flows and the radio/X-ray correlation. *Astron. Astrophys.* **414**, 895–903 (2004).
23. Turner, T. J., George, I. M., Nandra, K. & Turcan, D. On X-ray variability in Seyfert galaxies. *Astrophys. J.* **524**, 667–673 (1999).
24. Leighly, K. M. A comprehensive spectral and variability study of narrow-line Seyfert 1 galaxies observed by ASCA. I. Observations and time series analysis. *Astrophys. J., Suppl.* **125**, 297–316 (1999).
25. Bentz, M. C., Peterson, B. M., Pogge, R. W., Vestergaard, M. & Onken, C. A. The radius-luminosity relationship for active galactic nuclei: The effect of host-galaxy starlight on luminosity measurements. *Astrophys. J.* **644**, 133–142 (2006).
26. Peterson, B. M. *et al.* Central masses and broad-line region sizes of active galactic nuclei. II. A homogeneous analysis of a large reverberation-mapping database. *Astrophys. J.* **613**, 682–699 (2004).
27. Kollatschny, W., Zetzl, M. & Dietrich, M. Spectral line variability amplitudes in active galactic nuclei. *Astron. Astrophys.* **454**, 459–472 (2006).
28. Brandt, W. N., Boller, Th., Fabian, A. C. & Ruskowski, M. ROSAT High-Resolution Imager monitoring of extreme X-ray variability in the narrow-line quasar PHL 1092. *Mon. Not. R. Astron. Soc.* **303**, L53–L57 (1999).
29. Peterson, B. M. *et al.* X-ray and optical variability in NGC 4051 and the nature of narrow-line Seyfert 1 galaxies. *Astrophys. J.* **542**, 161–174 (2000).
30. Nagar, N. M., Oliva, E., Marconi, A. & Maiolino, R. NGC 5506 unmasked as a narrow line Seyfert 1: A direct view of the broad line region using near-IR spectroscopy. *Astron. Astrophys.* **391**, L21–L24 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the UK Particle Physics and Astronomy Research Council (PPARC). We thank P. Smith for discussions about statistics, P. Arevalo for determining the break timescale in GRS 1915+105 and D. Summons for providing some RXTE lightcurves of Cyg X-1.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to I.M.H. (imh@astro.soton.ac.uk).

Free fermion antibunching in a degenerate atomic Fermi gas released from an optical lattice

T. Rom¹, Th. Best¹, D. van Oosten¹, U. Schneider¹, S. Fölling¹, B. Paredes¹ & I. Bloch¹

Noise in a quantum system is fundamentally governed by the statistics and the many-body state of the underlying particles^{1–4}. The correlated noise^{5–7} observed for bosonic particles (for example, photons⁸ or bosonic neutral atoms^{9–14}) can be explained within a classical field description with fluctuating phases; however, the anticorrelations ('antibunching') observed in the detection of fermionic particles have no classical analogue. Observations of such fermionic antibunching are scarce and have been confined to electrons^{15–17} and neutrons¹⁸. Here we report the direct observation of antibunching of neutral fermionic atoms. By analysing the atomic shot noise^{3,10,19} in a set of standard absorption images of a gas of fermionic ⁴⁰K atoms released from an optical lattice, we find reduced correlations for distances related to the original spacing of the trapped atoms. The detection of such quantum statistical correlations has allowed us to characterize the ordering and temperature of the Fermi gas in the lattice. Moreover, our findings are an important step towards revealing fundamental fermionic many-body quantum phases in periodic potentials, which are at the focus of current research.

Bunching and antibunching effects in the detection of bosonic and fermionic particles are usually understood as a consequence of the constructive or destructive interference of the two possible propagation paths that two particles can follow to reach two detectors, as was first outlined in the famous experiments of Hanbury Brown and Twiss for the case of bosonic particles⁷. For the case of a fermionic atom cloud released from a periodic potential, the origin of spatial anticorrelations in the shot noise can be more naturally understood within the following picture. In the lowest energy band of a one-dimensional periodic potential, atoms can occupy Bloch states, each of them characterized by a crystal momentum $\hbar q$, as illustrated in Fig. 1, where q is the crystal wave vector and $\hbar$ denotes Planck's constant divided by 2π . Each Bloch state is a superposition of plane waves, which correspond to different real momenta $p_v = \hbar q + v2\hbar k$, where $k = 2\pi/\lambda$, λ being the wavelength of the laser light used to form the optical lattice potential, and v an integer number. When a particle with crystal momentum $\hbar q$ is released from the lattice potential, its wavefunction will expand freely, and after a time of flight (TOF) t , long enough to neglect the initial size of the system, it can be detected at any of the positions $x_v = (\hbar q + v2\hbar k)t/m$. These positions are equally spaced by $\ell = 2\hbar k t/m$, where m is the atomic mass. Conversely, if a particle is detected at any of the positions x_v , it had to emerge from a state with crystal momentum $\hbar q$. As the Pauli principle does not allow two fermionic particles to occupy the same Bloch state, the output of two detectors at positions x_v and $x_{v'}$ will thus be anticorrelated for any v, v' . That is, if one detector detects a particle, the second detector will not detect another particle. Therefore, if $n(x)$ represents the density of the expanding atom cloud detected at position x in a single run of the experiment, the spatial correlations $\langle n(x)n(x') \rangle$ will vanish for any $|x-x'|$ being an integer multiple of

ℓ . It is interesting to note that for bosonic particles, which can populate the same crystal momentum state, enhanced spatial correlations at these distances can arise, corresponding to a 'bunching' of the particles^{10,14}. For both bosons and fermions, the behaviour of $\langle n(x)n(x') \rangle$ for $|x-x'|$ different from these characteristic distances will depend on the many-body state of the trapped particles³.

The starting point for our experiment is an oblate-shaped degenerate Fermi gas in a crossed optical dipole trap with typically $2.0(5) \times 10^5$ atoms (the values in the brackets denote the 1σ uncertainty in the last shown digit) of ⁴⁰K at temperatures (T) as low as $T/T_F = 0.23(3)$, where T_F is the Fermi temperature. These atoms are transferred into a three-dimensional optical lattice potential with lattice depths of $V_x = V_y = 20E_r$ and $V_z = 10E_r$ in the x , y and z direction, respectively. Here, $E_r = \hbar^2 k^2 / 2m$ denotes the recoil energy. For our experimental parameters the atoms form a fermionic band insulator in the lowest energy band, which we verify using an adiabatic mapping technique^{20,21} (Fig. 2a inset). For the correlation

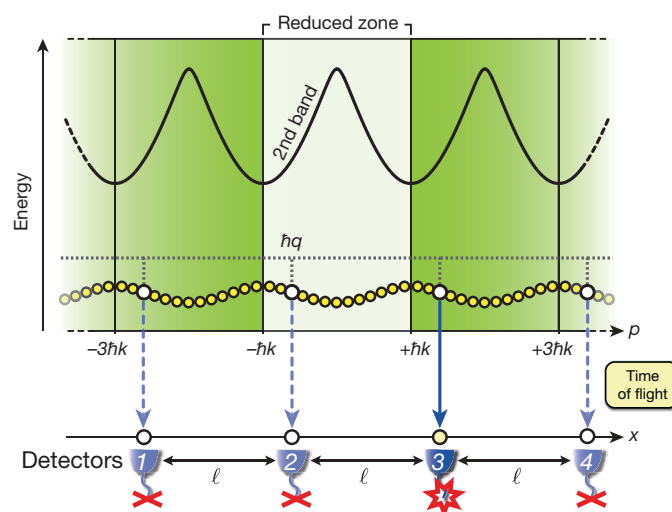


Figure 1 | Origin of anticorrelations in a Fermi gas released from an optical lattice. Each occupied Bloch state, labelled by a crystal momentum $\hbar q$, is represented by a dot in the reduced zone scheme (white region). The full occupation of the lowest energy band opposed to the empty second band describes the fermionic band insulating state. The periodically extended zone scheme (extension shown as green shaded region) shows that each Bloch state is a superposition of states with momenta equally spaced by $2\hbar k$. After the lattice is switched off abruptly, an atom with crystal momentum $\hbar q$ propagates freely during a time of flight t and can reach detectors equally spaced by a distance ℓ . If for example, detector number 3 detects a particle (yellow dot above detector), then owing to the single occupancy of each Bloch state dictated by the Pauli principle, detectors 1, 2 and 4 will not detect a particle (white dots above detectors).

¹Institut für Physik, Johannes Gutenberg-Universität, Staudingerweg 7, 55128 Mainz, Germany.

analysis, all optical potentials are suddenly switched off and images are recorded along the vertical direction (z) via standard absorption imaging after a TOF period of 10 ms.

Typically, 200 absorption images are taken for the same set of experimental parameters, yielding independent two-dimensional spatial density profiles, $n(\mathbf{x})$. An example of such a density profile and a corresponding cut through it can be seen in Fig. 2a, b. The atom cloud exhibits a gaussian-like envelope, as a result of the expansion of the Wannier function that characterizes the initial localization of the particles at each lattice site. Additionally, atomic noise is directly visible as fluctuations in the recorded single images. However, no distinct structure can be seen in this noise.

After discarding images showing technical artefacts ($\sim 30\%$ of the total images), a spatial correlation analysis is carried out. This yields the density–density correlation function:

$$C(\mathbf{d}) = \frac{\int \langle n(\mathbf{x} - \mathbf{d}/2) n(\mathbf{x} + \mathbf{d}/2) \rangle d^2x}{\int \langle n(\mathbf{x} - \mathbf{d}/2) \rangle \langle n(\mathbf{x} + \mathbf{d}/2) \rangle d^2x} \quad (1)$$

which gives the conditional probability of finding particles at two positions separated by a vector $\mathbf{d}$, averaged over the whole image. The brackets $\langle \cdot \rangle$ denote a statistical averaging over the set of images. Values of $C(\mathbf{d}) > 1$ correspond to bunching, whereas the characteristic antibunching for fermions shows up as $C(\mathbf{d}) < 1$.

Results of the above analysis are shown in Fig. 2c, d. Centred around the autocorrelation peak, a perfectly regular structure arises (Fig. 2c) in the correlation images. A horizontal cut through the centre of such a correlation image exhibits distinct dips where $C(\mathbf{d}_{\text{dip}}) < 1$ (Fig. 2d), demonstrating the antibunching of the atoms. We find that the spacing between these dips along the horizontal and vertical directions is in good agreement with the expected value of ℓ . In order to reliably detect and characterize these correlation dips, the signal to noise ratio in the resulting correlation image has to be sufficiently high. As the root mean square (r.m.s.) background noise decreases in good agreement with a $1/\sqrt{N_{\text{images}}}$ scaling, the signal to noise ratio can be successively increased by including more single shot absorption images (N_{images}) in the correlation analysis. When fitting two-dimensional gaussians to these dips, we find that at least

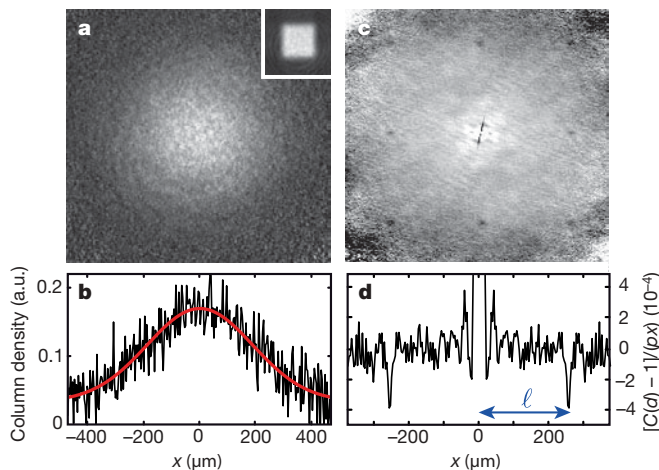


Figure 2 | Single shot absorption images and correlation analysis. **a**, Single absorption image of a fermionic ^{40}K atom cloud after 10 ms of free expansion. The inset shows a Brillouin zone mapping of the cloud, demonstrating that the Fermi gas is in a band insulating state. **b**, One-dimensional cut through **a** together with a gaussian fit (red). **c**, Spatial noise correlations obtained from an analysis of 158 independent images, showing an array of eight dark dots. **d**, A horizontal profile through the centre of the image shows that these dots correspond to correlation dips, with the characteristic spacing ℓ . The profile has been high-pass filtered to suppress a broad gaussian background that we attribute to shot to shot fluctuations in the atom number.

50 images are needed for a reliable fit, if width, position and amplitude of the dips are free parameters.

In order to experimentally investigate the temperature dependence of the measured correlation signal, we have prepared Fermi gases at different temperatures by varying the end point of the forced evaporation procedure in the optical dipole trap. Before the lattice potentials are ramped up, the optical trap depth is adjusted to obtain a defined external confinement, equal for all temperatures. For each temperature, a correlation analysis for a series of absorption images is performed. As Fig. 3 shows, we find a distinct decrease of the correlation amplitude with increasing temperature. Antibunching is, however, still clearly visible up to temperatures of $T/T_F \approx 1$. Note that all quoted temperatures refer to the initial temperatures in the optical dipole trap without the lattice, as we currently do not have an independent means to measure the temperature of the quantum gas in the lattice.

In order to understand the strength of the correlation signals and specifically their temperature dependence, we consider a one-dimensional system of N fermionic atoms in a lattice in the presence of an external harmonic confinement. We assume that the lattice is deep enough to suppress tunnelling processes and that the temperature is low enough such that the atoms only occupy the lowest on-site vibrational state. Under these conditions, the occupation of the lattice sites is given by a Fermi–Dirac distribution of the form:

$$f_j = \frac{1}{1 + \exp((\frac{1}{2} m \omega^2 (\lambda/2)^2 j^2 - \mu) / k_B T^{\text{lat}})} \quad (2)$$

where the integer number j labels the lattice site, T^{lat} denotes the temperature in the lattice, and k_B is Boltzmann's constant. The trapping frequency of the external harmonic confinement is denoted by ω , and the chemical potential μ is defined through $\sum_j f_j = N$. At zero temperature, starting from the centre of the trap, the atoms fill the lattice sites one-by-one, forming a Fermi sea in real space (Fig. 4 left inset). In analogy to the Fermi momentum of free fermions, we define a Fermi radius $R_F = N\lambda/4$, which characterizes the extension of the atomic cloud. The Fermi temperature in the lattice, T_F^{lat} , can then be written as $k_B T_F^{\text{lat}} = 1/2 m \omega^2 R_F^2$. As temperature increases, atoms can access lattice sites further out in the trap, leading to an increase in the spatial extension of the atom cloud.

A similar calculation to the one presented in ref. 10 yields the fermionic noise correlator:

$$C(d) = 1 - \frac{1}{N^2} \left| \sum_j f_j e^{i2\pi d j / \ell} \right|^2 \quad (3)$$

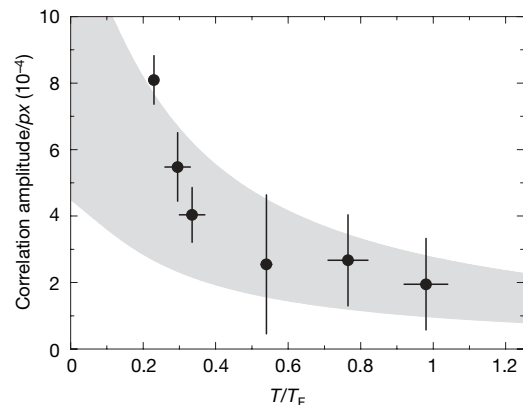


Figure 3 | Measured correlation amplitude versus temperature of the atomic cloud before loading into the optical lattice. For each data point, two to four independent runs with about 200 images in total were averaged (data show mean $\pm$ s.d.). The shaded area indicates the expected values of the correlation amplitude for the two-dimensional version of the theory discussed in the text (see Methods), taking into account our experimental uncertainties in the atom number.

where the minus sign on the right-hand side causes dips at integer multiples of ℓ . For zero temperature, the above equation takes the simple form $C(d) = 1 - \left[\frac{\sin(\pi d / \ell N)}{N \sin(\pi d / \ell)} \right]^2$. The temperature dependence of the shape of one of these dips is shown in Fig. 4. The amplitude of the dip is always equal to 1, regardless of temperature. However, its width decreases with increasing temperature. This leads to a decrease in the signal strength of the dip, which we characterize by the spatial integral of the correlation function (see equation (3)) over a distance ℓ . We can understand these features in the following way. The amplitude of 1, corresponding to the vanishing of the correlation function $C(d)$, is a universal feature for fermions. It reflects the fact that two fermions cannot occupy the same crystal momentum state. On the other hand, the detailed shape of the dip directly reflects the density–density correlations in crystal momentum space of the fermionic system before expansion. These correlations can be characterized by a correlation length in momentum space, which determines the width of the dip. For non-interacting fermions, this correlation length is inversely proportional to the size of the system. This explains the observation that when temperature increases, leading to an expansion of the system in the trap, the width of the dip decreases. It is interesting to note that for the case of zero temperature, the wings of the dip oscillate with a period inversely proportional to the Fermi radius. These oscillations are analogous to the Friedel oscillations that a system of free fermions exhibits in real space.

Let us compare these findings to our experimental results. For a typical system size of $N = 100$ in the x and y dimension, we expect the width of each dip to be of the order of $1 \mu\text{m}$ on the basis of the above model, which is much smaller than the actual optical resolution of our imaging system of $5.6 \mu\text{m}$. Therefore, the observed shape and width of our correlation dips are predominantly determined by the point spread function of our optical imaging system (approximately a gaussian) and do not allow us to observe the detailed shape of the correlation dips predicted by theory. However, as the point spread function leaves the volume of a correlation dip constant, a change in the width of the theoretically predicted signal directly translates into a change of the observed correlation amplitude. For our lowest temperatures, we observe a correlation amplitude of $8(1) \times 10^{-4}$, in good agreement with the expected value at zero temperature of $9(4) \times 10^{-4}$ for our experimental parameters, taking additionally into account the integration along the line of sight¹⁰. For increasing temperatures, we have compared the observed decrease in the correlation signal with the one predicted by a two-dimensional version

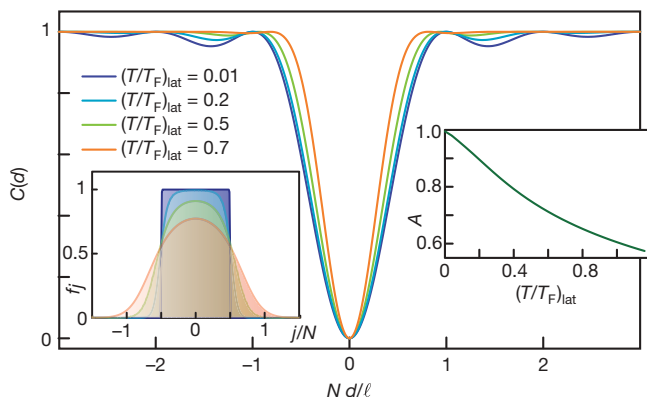


Figure 4 | Calculated noise correlations for ^{40}K atoms in a one-dimensional lattice. In the main panel, the different curves correspond to different values of the temperature $(T/T_F)_{\text{lat}}$ in the lattice. The left inset shows the corresponding spatial density profiles of the system for different values of $(T/T_F)_{\text{lat}}$. The right inset shows the area A under the correlation dip as a function of $(T/T_F)_{\text{lat}}$. The parameters used were $N = 100$ and an external confinement of $\omega = 2\pi \times 40 \text{ Hz}$, yielding $T_F^{\text{lat}} \approx 350 \text{ nK}$.

of the model presented above (Methods). Assuming adiabatic heating when the atoms are loaded into the lattice potential²², we find good quantitative agreement with theory (Fig. 3). These results show that the correlation signal can be used as an efficient thermometer for fermions in an optical lattice.

In conclusion, we have demonstrated antibunching of free neutral fermionic atoms using a degenerate single component Fermi gas released from a three dimensional optical lattice potential. Recently it has come to our attention that fermionic antibunching has also been observed in an ultracold gas of metastable ^3He atoms²³. In our experiment the anticorrelations obtained from a shot noise analysis of standard absorption images have allowed us to reveal the quantum statistics and furthermore identify the ordering and temperature of the atoms in the periodic potential up to a temperature of $T/T_F \approx 1$. These measurements show that noise correlations are a robust tool for further studies of degenerate Fermi systems. In particular, this method could allow the unambiguous detection of fermionic quantum phases, such as an antiferromagnetically ordered Néel phase²⁴, which is strongly connected to models of high- T_c superconductivity^{25,26}. Further complex orders could be revealed by applying this method to low-dimensional quantum systems^{4,27}, mixtures^{28,29}, quantum phases with disorder³⁰ or supersolid phases³¹.

METHODS

Experimental set-up and cooling cycle. In the experiment, we sympathetically cool ^{40}K with ^{87}Rb in a two step process. First, both species are trapped in a magnetic trap of the quadrupole Ioffe type, with ^{87}Rb in the $|F=2, m_F=2\rangle$ and ^{40}K in the $|F=9/2, m_F=9/2\rangle$ substates. Radio-frequency (r.f.)-evaporative cooling of ^{87}Rb is used to cool the mixture to a temperature of $2 \mu\text{K}$. Both species are then transferred into a crossed optical dipole trap formed at the intersection of two orthogonal elliptical gaussian laser beams (waists $w_{x,y} = 150 \mu\text{m}$, $w_z = 40 \mu\text{m}$) at a wavelength of $1,030 \text{ nm}$. After switching off the magnetic trap, a constant offset field of 13.6 G is applied. Then ^{87}Rb is transferred to the $|F=1, m_F=1\rangle$ state with an adiabatic microwave sweep, and ^{40}K is transferred to the $|F=9/2, m_F=-9/2\rangle$ state with an adiabatic r.f. sweep across the magnetic substates. The depth of the optical dipole trap is subsequently lowered to further evaporate the two species. At the end of this cooling process, we are left with up to $2.2(5) \times 10^5$ atoms of ^{40}K , and $2.0(5) \times 10^5$ atoms of ^{87}Rb , in the optical trap. After removing the rubidium cloud completely via a resonant laser pulse, we are left with a single component Fermi gas of typically $2.0(5) \times 10^5$ atoms at an initial temperature T as low as $T/T_F \approx 0.23(3)$.

Optical lattices. The cold fermionic quantum gas is loaded into a three-dimensional optical lattice potential formed by three mutually orthogonal optical standing waves. The laser beams used for the optical standing waves have waists of $150 \mu\text{m}$ at the position of the trapped quantum gas and a wavelength of $\lambda = 755 \text{ nm}$, blue detuned to the atomic D_1 and D_2 transitions of ^{40}K . The lattice potentials are successively ramped up to final lattice depths of $V_x = V_y = 20 E_r$, and $V_z = 10 E_r$. This is done in an 'S'-shaped ramp within 40 ms , first in the x and y directions and finally in the z direction.

Determining correlation signal amplitudes. Correlation signal amplitudes are extracted from the unprocessed correlation images by fitting the dips of the correlation images with an inverted two dimensional gaussian. First, the positions and widths of the anticorrelation dips are determined from three correlation images with $T/T_F = 0.23(3)$. In subsequent fits to all images, the dip amplitudes were obtained using these widths and positions. The correlation signal amplitude of an image is given as the average over the four dips at positions corresponding to the principal reciprocal lattice vectors. Note that the central autocorrelation peak shows a substructure due to technical artefacts.

Temperature dependence in two dimensions. A generalization of the theory model described above to the two-dimensional case is straightforward. The correlation function takes then the form $C(\mathbf{d}) = 1 - \frac{1}{N^2} \left| \sum_j f_j e^{i2\pi \mathbf{j} \cdot \mathbf{d} / \ell} \right|^2$, where $\mathbf{j} = (j_x, j_y)$, with j_x, j_y being integer numbers. The integral of the function $C(\mathbf{d})$ over a $\ell \times \ell$ cell, which characterizes the signal strength of the dips, can be calculated to be:

$$\frac{\ell^2}{N^2} \sum_j f_j^2 = \frac{\ell^2}{N} \left(1 - \frac{T}{T_F^{\text{lat}}} \left(1 - e^{-T_F^{\text{lat}}/T} \right) \right) \quad (4)$$

where $k_B T_F^{\text{lat}} = m\omega^2(\lambda/2)^2 N/2\pi$ is the Fermi temperature of the atoms in the two-dimensional lattice. This result has been used to determine the region of amplitudes that, taking into account the uncertainty in the number of particles

and the heating of the atoms when loading the optical lattice, is expected for the experimental dips.

Received 15 August; accepted 6 October 2006.

- Blanter, Y. M. & Büttiker, M. Shot noise in mesoscopic conductors. *Phys. Rep.* **336**, 1–166 (2000).
- Beenakker, C. W. J. & Schönberger, C. Quantum shot noise. *Phys. Today* **56**, 37–42 (2003).
- Altman, E., Demler, E. & Lukin, M. D. Probing many-body states of ultracold atoms via noise correlations. *Phys. Rev. A* **70**, 013603 (2004).
- Polkovnikov, A., Altman, E. & Demler, E. Interference between independent fluctuating condensates. *Proc. Natl Acad. Sci. USA* **103**, 6125–6129 (2006).
- Brown, R. H. & Twiss, R. Q. Correlation between photons in two coherent beams of light. *Nature* **177**, 27–29 (1956).
- Brown, R. H. & Twiss, R. Q. The question of correlation between photons in coherent light rays. *Nature* **178**, 1447–1448 (1956).
- Baym, G. The physics of Hanbury Brown-Twiss intensity interferometry: From stars to nuclear collisions. *Acta Phys. Pol. B* **29**, 1839–1884 (1998).
- Mandel, L. & Wolf, E. *Optical Coherence and Quantum Optics* (Cambridge Univ. Press, Cambridge, New York, 1995).
- Yasuda, M. & Shimizu, F. Observation of two-atom correlation of an ultracold neon atomic beam. *Phys. Rev. Lett.* **77**, 3090–3093 (1996).
- Fölling, S. *et al.* Spatial quantum noise interferometry in expanding ultracold atom clouds. *Nature* **434**, 481–484 (2005).
- Schellekens, M. *et al.* Hanbury Brown Twiss effect for ultracold quantum gases. *Science* **310**, 648–651 (2005).
- Öttl, A., Ritter, S., Köhl, M. & Esslinger, T. Correlations and counting statistics of an atom laser. *Phys. Rev. Lett.* **95**, 090404 (2005).
- Esteve, J. *et al.* Observations of density fluctuations in an elongated Bose gas: Ideal gas and quasicondensate regimes. *Phys. Rev. Lett.* **96**, 130403 (2006).
- Spielman, I. B., Phillips, W. D. & Porto, J. V. The Mott insulator transition in two dimensions. Preprint at (<http://arxiv.org/cond-mat/0606216>) (2006).
- Oliver, W. D., Kim, J., Liu, R. C. & Yamamoto, Y. Hanbury Brown and Twiss-type experiment with electrons. *Science* **284**, 299–301 (1999).
- Henny, M. *et al.* The fermionic Hanbury Brown and Twiss experiment. *Science* **284**, 296–298 (1999).
- Kiesel, H., Renz, A. & Hasselbach, F. Observation of Hanbury Brown-Twiss anticorrelations for free electrons. *Nature* **418**, 392–394 (2002).
- Iannuzzi, M., Orecchini, A., Sacchetti, F., Facchi, P. & Pascazio, S. Direct experimental evidence of free-fermion antibunching. *Phys. Rev. Lett.* **96**, 080402 (2006).
- Greiner, M., Regal, C. A., Stewart, J. T. & Jin, D. S. Probing pair-correlated fermionic atoms through correlations in atom shot noise. *Phys. Rev. Lett.* **94**, 110401 (2005).
- Greiner, M., Bloch, I., Mandel, O., Hänsch, T. W. & Esslinger, T. Exploring phase coherence in a 2D lattice of Bose-Einstein condensates. *Phys. Rev. Lett.* **87**, 160405 (2001).
- Köhl, M., Moritz, H., Stöferle, T., Günter, K. & Esslinger, T. Fermionic atoms in a three dimensional optical lattice: Observing Fermi-surfaces, dynamics and interactions. *Phys. Rev. Lett.* **94**, 080403 (2005).
- Köhl, M. Thermometry of fermionic atoms in an optical lattice. *Phys. Rev. A* **73**, 031601 (2006).
- Westbrook, C. I. *et al.* Producing and detecting correlated atoms. Preprint at (<http://arxiv.org/quant-ph/0609019>) (2006).
- Werner, F., Parcollet, O., Georges, A. & Hassan, S. R. Interaction-induced adiabatic cooling and antiferromagnetism of cold fermions in optical lattices. *Phys. Rev. Lett.* **95**, 056401 (2005).
- Lee, P. A., Nagaosa, N. & Wen, X.-G. Doping a Mott insulator: Physics of high-temperature superconductivity. *Rev. Mod. Phys.* **78**, 17–85 (2006).
- Hofstetter, W., Cirac, J. I., Zoller, P., Demler, E. & Lukin, M. D. High-temperature superfluidity of fermionic atoms in optical lattices. *Phys. Rev. Lett.* **89**, 220407 (2002).
- Mathey, L., Altman, E. & Vishwanath, A. Noise correlations in one-dimensional systems of ultra-cold fermions. Preprint at (<http://arxiv.org/cond-mat/0507108>) (2005).
- Wang, D.-W., Lukin, M. D. & Demler, E. Engineering superfluidity in Bose-Fermi mixtures of ultracold atoms. *Phys. Rev. A* **72**, 051604 (2005).
- Ahufinger, V., Sanchez-Palencia, L., Kantian, A., Sanpera, A. & Lewenstein, M. Disordered ultracold atomic gases in optical lattices: A case study of Fermi-Bose mixtures. *Phys. Rev. A* **72**, 063616 (2005).
- Rey, A. M., Indubala, I. S. & Clark, C. W. Quantum coherence of hard-core bosons and fermions: Extended, glassy and Mott phases. *Phys. Rev. A* **73**, 063610 (2006).
- Scarola, V. W., Demler, E. & Sarma, S. D. Searching for a supersolid in cold-atom optical lattices. *Phys. Rev. A* **73**, 051601(R) (2006).

Acknowledgements This work was supported by the DFG, and by the EU under a Marie-Curie excellence grant (QUASICOMBS) and an Integrated Project (SCALA). We acknowledge the technical assistance of T. Berg in the construction of the apparatus.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to I.B. (bloch@uni-mainz.de).

Controlled injection and acceleration of electrons in plasma wakefields by colliding laser pulses

J. Faure¹, C. Rechatin¹, A. Norlin¹, A. Lifschitz¹, Y. Glinec¹ & V. Malka¹

In laser-plasma-based accelerators¹, an intense laser pulse drives a large electric field (the wakefield) which accelerates particles to high energies in distances much shorter than in conventional accelerators. These high acceleration gradients, of a few hundreds of gigavolts per metre, hold the promise of compact high-energy particle accelerators. Recently, several experiments have shown that laser-plasma accelerators can produce high-quality electron beams, with quasi-monoenergetic energy distributions at the 100 MeV level^{2–4}. However, these beams do not have the stability and reproducibility that are required for applications. This is because the mechanism responsible for injecting electrons into the wakefield is based on highly nonlinear phenomena⁵, and is therefore hard to control. Here we demonstrate that the injection and subsequent acceleration of electrons can be controlled by using a second laser pulse⁶. The collision of the two laser pulses provides a pre-acceleration stage which provokes the injection of electrons into the wakefield. The experimental results show that the electron beams obtained in this manner are collimated (5 mrad divergence), monoenergetic (with energy spread <10 per cent), tuneable (between 15 and 250 MeV) and, most importantly, stable. In addition, the experimental observations are compatible with electron bunch durations shorter than 10 fs. We anticipate that this stable and compact electron source will have a strong impact on applications requiring short bunches, such as the femtotoxicity of water⁷, or high stability, such as radiotherapy with high-energy electrons^{8,9} or radiography¹⁰ for materials science.

In laser-plasma-based accelerators, the longitudinal accelerating electric field is excited via the ponderomotive force of an ultrashort and ultra-intense laser. This force is proportional to the gradient of the laser intensity I . The normalized laser vector potential is defined as a_0 , with $a_0^2 \propto I$. The ponderomotive force pushes electrons outward and separates them from the ions, thus creating a travelling electric field, with a phase velocity v_p close to the speed of light in vacuum c , as is required for accelerating particles to relativistic energies. This wakefield can have a longitudinal electric field $E_z > 100 \text{ GV m}^{-1}$ and the characteristic scale length of the accelerating structure is the plasma wavelength $\lambda_p = 10\text{--}30 \mu\text{m}$, for typical plasma densities of $n_e = 10^{18}\text{--}10^{19} \text{ cm}^{-3}$. To be trapped and accelerated, electrons need to be injected into the wakefield with sufficient initial energy.

The experiments reported^{2–4} in 2004 operated in the bubble regime⁵ (or blow-out regime¹¹), in which electrons were ‘self-injected’ into the wakefield, allowing the trapping and the acceleration of quasi-monoenergetic bunches. In these experiments, a single laser pulse was responsible for trapping and accelerating electrons. This physics is highly nonlinear and can be explained as follows: the nonlinear evolution of the laser pulse via self-focusing¹² and self-compression¹³ leads to an increase of the laser intensity and to the formation of an electron-evacuated cavity (the bubble), filled with

ions and surrounded by a dense wall of electrons. When the electron density at the walls reaches a threshold value, self-injection occurs at the back of the bubble. Injection stops when the charge density of the trapped bunch is comparable to the charge density at the bubble walls. This short and localized injection leads to the formation of a quasi-monoenergetic electron bunch. However, this self-injection mechanism depends crucially on the nonlinear evolution of the laser pulse, which might be different from shot to shot. This causes injection to occur at different times, and our previous experiments⁴ have suffered large fluctuations in the self-injected electron beam in terms of electron energy distribution: some shots produce electron beams with a well-defined monoenergetic peak (one out of 3–5 shots), whereas others produce beams with several peaks or sometimes even a plateau distribution. Recently, other groups have obtained more stable beams using tailored gas jet targets¹⁴ or capillary discharges¹⁵.

It is believed that the stabilization of the electron beam can be achieved by controlling the injection of electrons. Therefore, a method of injecting electrons ‘externally’ is highly desirable. In addition, the production of monoenergetic electron bunches requires that the injected beam load have a duration shorter than $\lambda_p/c = 30\text{--}100 \text{ fs}$. However, the production of such short bunches is at the edge of conventional radio-frequency (r.f.) accelerator technology and no successful experiment involving an external r.f. injector has been demonstrated so far. In contrast, our experimental results demonstrate a method that greatly enhances the stability of the electron beam by controlling the injection of electrons in the wakefield using a second laser pulse. The idea of using an additional laser pulse for injecting electrons was first proposed in ref. 16. It was further developed⁶ into a method based on counter-propagating laser pulses. In its simplest form, this method uses two counter-propagating ultrashort laser pulses with the same central wavelength and polarization¹⁷. The first laser pulse—the ‘pump’ pulse (with laser strength a_0)—creates a wakefield, whereas the second laser—the ‘injection’ pulse (with laser strength $a_1 < a_0$)—will only be used for injecting electrons. The laser pulses collide in the plasma and their interference creates a laser beatwave pattern with phase velocity $v_{bw} \approx 0$.

The beatwave pattern is a standing wave, with characteristic spatial scale $\lambda_0/2$, where λ_0 is the laser central wavelength. Because of this small scale length, the ponderomotive force of the beatwave is very large: $F_{bw} \propto 2a_0a_1/\lambda_0$. Owing to its zero phase velocity and large ponderomotive force, the beatwave can pre-accelerate plasma background electrons. Under certain conditions for a_0 and a_1 , this pre-acceleration permits the injection and trapping of electrons in the wakefield and further acceleration to relativistic energies. Analytical work¹⁷ and simulations^{17,18} have shown that this two-stage acceleration mechanism can lead to the production of high-quality electron bunches, with narrow energy spread, small divergence and ultrashort duration, even when using relatively modest lasers ($a_0 = 1$, $a_1 = 0.3$).

¹Laboratoire d’Optique Appliquée, ENSTA, CNRS, Ecole Polytechnique, UMR 7639, 91761 Palaiseau, France.

In the experiment, we focused two ultrashort 30 fs laser pulses with linear polarization at the edge of a 2 mm supersonic helium gas jet¹⁹ (see Supplementary Information). The gas was rapidly ionized by the front of the laser pulses and provided the plasma medium. The pump pulse was focused to an intensity of $I_0 = 3.4 \times 10^{18} \text{ W cm}^{-2}$, for which $a_0 = 1.3$. The injection pulse was collinear and counter-propagating, with an intensity of $I_1 = 4 \times 10^{17} \text{ W cm}^{-2}$, for which $a_1 = 0.4$. The electron beam was measured with an electron spectrometer²⁰, which gives access to the electron beam angular distribution, energy distribution and charge. Figure 1a–c shows the electron distribution obtained when the pump pulse alone is fired into the gas jet: this is the usual self-injected beam obtained in the bubble regime. Achieving controlled injection requires that self-injection be turned off. Figure 1a–c shows that by decreasing the plasma density from $n_e = 1.25 \times 10^{19} \text{ cm}^{-3}$ to $n_e = 7.5 \times 10^{18} \text{ cm}^{-3}$, we were able to suppress the self-injected electron beam by operating below the threshold of self-injection. For densities lower than $n_e = 7.5 \times 10^{18} \text{ cm}^{-3}$, the nonlinear evolution of the laser pulse through self-focusing and self-compression was not strong enough to cause significant injection of electrons into the wakefield. However, at $n_e = 7.5 \times 10^{18} \text{ cm}^{-3}$, the addition of the injection pulse produced a monoenergetic electron beam around 200 MeV, as shown in Fig. 1d. In the latter, the two laser beams had parallel polarization, allowing them to interfere and set up the beatwave pattern necessary for injecting electrons. When the polarizations of the two beams are orthogonal, no electron beam is produced; see Fig. 1e.

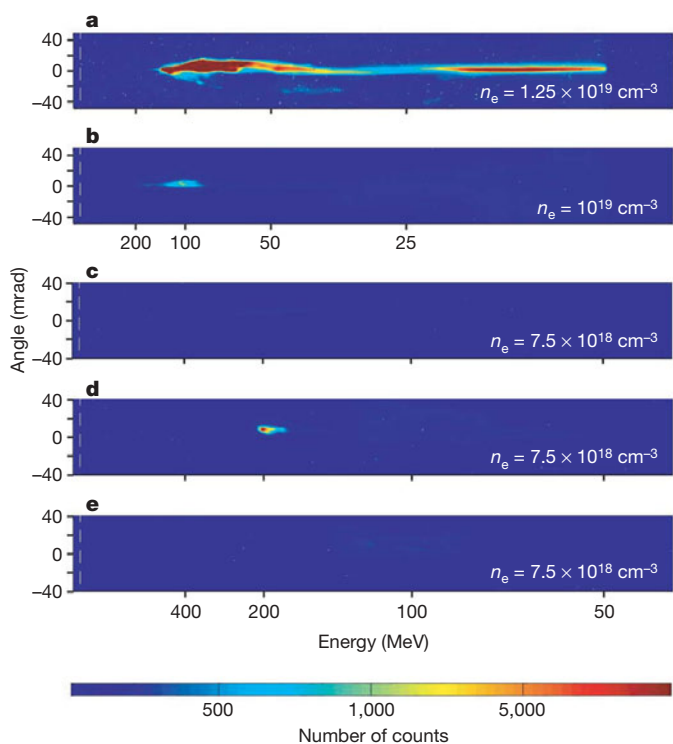


Figure 1 | Raw images of the electron beam obtained with the electron spectrometer. Horizontal axis, electron energy; vertical axis, angular divergence. The colour scale reflects the number of counts which gives an indication of the beam charge. **a–c** were obtained with the pump laser pulse only. **a**, The image shows an intense self-injected electron beam with a broad energy distribution ($n_e = 1.25 \times 10^{19} \text{ cm}^{-3}$). In **b** the self-injected electron beam has less charge but a quasi-monoenergetic distribution ($n_e = 10^{19} \text{ cm}^{-3}$). In **c** there is no electron beam, because the density is below the threshold for self-injection ($n_e = 7.5 \times 10^{18} \text{ cm}^{-3}$). **d** was obtained by colliding the pump with the injection pulse with parallel polarizations, at the same plasma density ($n_e = 7.5 \times 10^{18} \text{ cm}^{-3}$). A high-quality monoenergetic electron beam at 200 MeV is produced. **e**, When the polarizations of the laser beams are crossed, no injection occurs.

Table 1 | Statistics of the electron beam parameters over 20 shots

Peak energy (mean $\pm$ s.d.)	117 $\pm$ 7 MeV
Energy spread FWHM (mean $\pm$ s.d.)	11 $\pm$ 2 %
Charge (mean $\pm$ s.d.)	19 $\pm$ 6.8 pC
Beam divergence FWHM (mean $\pm$ s.d.)	5.8 $\pm$ 2 mrad
Beam pointing stability (mean $\pm$ s.d.)	0 $\pm$ 1.8 mrad

This laser polarization test strongly supports the idea that electron injection comes from the beatwave at the collision of the two lasers. It also suggests both that the injected charge can be tuned simply by rotating the polarization of one of the laser beams, and that the electrons are injected during the overlap of the two laser pulses (~ 30 fs). Thus, the trapped electron bunch originates from a single period of the plasma wave (because $\lambda_p/c = 40 \text{ fs} > 30 \text{ fs}$). Although some electrons might slip backwards in the following arches of the plasma wave, the measured monoenergetic bunch has a high energy, with a small energy spread and divergence, indicating that it resides in an accelerating and focusing phase of the wakefield, which has a length of $\lambda_p/4$ (see Supplementary Information for details). These considerations suggest that our observations are compatible with electron bunch durations shorter than $\tau < \lambda_p/4c = 10 \text{ fs}$.

One of the main differences between the colliding pulse injection scheme and the self-injected regime is that the electron beams obtained in this manner are very stable: every shot consistently gives the same monoenergetic electron distribution. A series of 20 consecutive shots was carried out to estimate the statistical fluctuations of the beam. Figure 2 shows a typical electron spectrum from this series, whereas Table 1 summarizes the mean parameters of the beam and their standard deviation.

In addition to enhanced stability, tuning the electron beam energy can be achieved by adjusting the position of the collision in the gas jet. The collision point can be modified by simply changing the delay between the two laser pulses. If the lasers collide at the entrance of the gas jet, electrons will be injected at an early stage and they can be accelerated over the whole gas jet length (2 mm). Thus their energy will be high. On the contrary, injection at the exit of the gas jet will limit the acceleration length and will lead to a low-energy beam.

This is demonstrated in Fig. 3, which shows the evolution of the beam peak energy (red curve) from about 50 MeV (the limit of the spectrometer) to 250 MeV. The small size of the error bars illustrates the stability of the beam energy. The blue curve shows how the energy spread $\delta E/E$ decreases from 20% to 5% (the resolution of the spectrometer) with beam energy. This is because the width of the monoenergetic bunch (at full-width at half-maximum, FWHM), $\delta E = 10$ –20 MeV, stays constant over the acceleration length. The beam charge is in the tens of picoCoulomb range: it goes as high as 60–80 pC below 100 MeV, it is in the 15–30 pC range at 100–200 MeV and drops below 10 pC when the beam energy is higher than 200 MeV. In

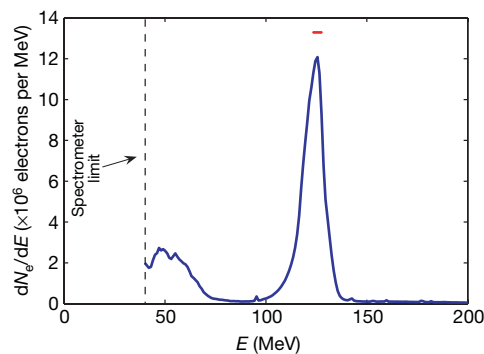


Figure 2 | A typical quasi-monoenergetic electron spectrum obtained by colliding pulse injection. Here, injection occurred at the middle of the gas jet, at position $z_{\text{inj}} = 0$. The beam charge is 25 pC, the peak energy is 125 MeV and the energy spread is 9%. The red error bar corresponds to the resolution of the spectrometer, which is 4% in this case.

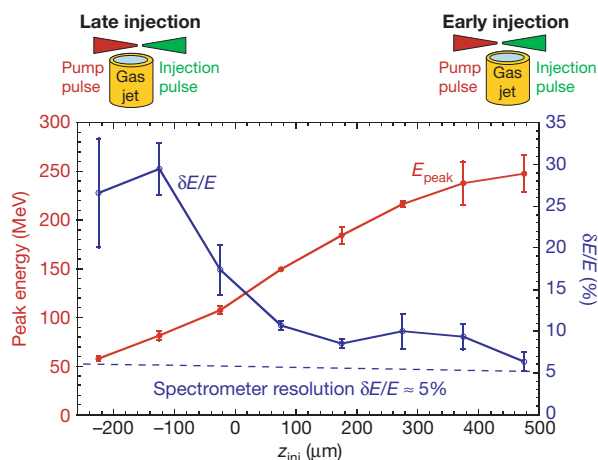


Figure 3 | Evolution of the electron beam peak energy and its energy spread. z_{inj} is the injection position, that is, the position at which the two laser pulses collide. The electron beam peak energy is shown in red, and the energy spread in blue. Each point is an average of 3–5 shots and the error bars correspond to the standard deviation. The position $z_{\text{inj}} = 0$ corresponds to injection at the middle of the gas jet, whereas $z_{\text{inj}} = 500 \mu\text{m}$ corresponds to early injection close to the entrance of the gas jet, as shown by the schematics on top of the graph.

addition, we can estimate the value of the average accelerating electric field from Fig. 3: the electron beam gains 190 MeV over 700 μm , which gives $E_z = 270 \text{ GV m}^{-1}$. During other scans and using a smaller magnet in the electron spectrometer, we have produced monoenergetic beams as low as 15 MeV. Thus the beam energy can be tuned simply by changing the position of the collision, bringing enormous flexibility to this accelerator.

The theory of colliding pulse injection^{6,17} was developed in the linear regime where $a_0 \ll 1$. Although our parameters are beyond the validity of the theory ($a_0 = 1.3$), we found it helpful to compare our experimental results to theoretical estimates. Following ref. 17, calculations show that for $a_0 = 1.3$, the trapped charge for $a_1 = 0.4$ is 31 pC, the beam energy reaches 70 MeV after 0.9 mm and the bunch duration after acceleration is 5 fs. The energy spread is $\delta E = 12 \text{ MeV}$ and the divergence (FWHM) is 6 mrad. These values are reasonably close to those found experimentally and show that analytical models of laser-triggered injection^{6,17} give a qualitative understanding of the physics at play (see Supplementary Information).

Thus by demonstrating a new method of electron injection into plasma wakefields, we have produced a high-quality electron source with a high level of stability. Quality, stability and control of beam parameters make these new laser-plasma accelerators attractive candidates for applications in various fields. We anticipate that increasing the beam charge to the nanoCoulomb level and the electron energy to several gigaelectronvolts using waveguides¹⁵ or more powerful lasers will aid the development of compact X-ray femto-second radiation sources.

METHODS

Laser pulses. The two laser pulses originated from a 10 Hz titanium-doped sapphire, chirped pulse amplification laser system. The pulses had linear polarization and a central wavelength of 820 nm. The pump pulse was focused using an $f = 1 \text{ m}$ on-axis spherical mirror. The parameters were: pulse energy 720 mJ, pulse duration $30 \pm 2 \text{ fs}$ at FWHM. The focal spot was carefully measured with a 12-bit charge-coupled device (CCD) camera; it was slightly elliptical: $16 \times 21 \mu\text{m}$ at FWHM. This gave a peak intensity of $3.4 \times 10^{18} \text{ W cm}^{-2}$ and $a_0 = 1.3$. The injection beam was focused with an $f = 1 \text{ m}$ off-axis parabola. The parameters were 250 mJ, $30 \pm 2 \text{ fs}$ FWHM, focal spot size $31 \mu\text{m}$ at FWHM, peak intensity $4 \times 10^{17} \text{ W cm}^{-2}$, and $a_1 = 0.4$. A third low-intensity laser pulse was used for synchronizing and overlapping the two main pulses by means of side-view imaging and shadowgraphy. Finally, a Faraday rotator was inserted in the laser system to prevent laser feedback from the experiment.

Gas jet. The gas jet was a supersonic helium gas jet with a 2 mm diameter. The lasers were focused 800 μm above the opening of the nozzle. The density profile was carefully characterized at this position: it had a 1.5 mm flat plateau surrounded by 400 μm gradients on each side. Both lasers were focused at the right edge of the plateau.

Electron spectrometer. The electron spectrometer consisted of a LANEX phosphor screen and a bending magnet (more details can be found in refs. 4 and 20). A round pole electromagnet providing an effective magnetic field of $B_{\text{eff}} = 1.6 \text{ T}$ over 2.5 cm was initially used (Fig. 1a and b were obtained with this magnet). In this configuration, electrons with energies greater than 10 MeV could be measured and the resolution at 200 MeV was 15%. Later in the experiment, a rectangular permanent magnet with $B_{\text{eff}} = 1.1 \text{ T}$ over 10 cm was used in order to increase the resolution. In this case, electrons with energies greater than 45 MeV could be measured, with a resolution of 5% at 200 MeV. The charge of the electron beam was obtained by measuring the number of photons emitted by the phosphor screen. The emission of the phosphor screen was independently calibrated using a 10 MeV r.f. accelerator.

Received 4 July; accepted 26 October 2006.

1. Tajima, T. & Dawson, J. M. Laser electron accelerator. *Phys. Rev. Lett.* **43**, 267–270 (1979).
2. Mangles, S. *et al.* Mono-energetic beams of relativistic electrons from intense laser-plasma interactions. *Nature* **431**, 535–538 (2004).
3. Geddes, C. G. R. High-quality electron beams from a laser wakefield accelerator using plasma-channel guiding. *Nature* **431**, 538–541 (2004).
4. Faure, J. *et al.* A laser-plasma accelerator producing monoenergetic electron beams. *Nature* **431**, 541–544 (2004).
5. Pukhov, A. & Meyer-ter-Vehn, J. Laser wake field acceleration: the highly nonlinear broken-wave regime. *Appl. Phys. B* **74**, 355–361 (2002).
6. Esarey, E., Hubbard, R. F., Leemans, W. P., Ting, A. & Sprangle, P. Electron injection into plasma wake fields by colliding laser pulses. *Phys. Rev. Lett.* **79**, 2682–2685 (1997).
7. Brozek-Pluska, B., Gliger, D., Hallou, A., Malka, V. & Gauduel, Y. A. Direct observation of elementary radical events: low- and high-energy radiation femtochemistry in aqueous solutions. *Radiat. Chem.* **72**, 149–159 (2005).
8. Desrosiers, C., Moskvina, V., Bielajew, A. F. & Papiez, L. 150–250 MeV electron beams in radiation therapy. *Phys. Med. Biol.* **45**, 1781–1805 (2000).
9. Glinec, Y. *et al.* Radiotherapy with laser-plasma accelerators: Monte-Carlo simulation of dose deposited by an experimental quasi-monoenergetic electron beam. *Med. Phys.* **33**, 155–162 (2006).
10. Glinec, Y. *et al.* High resolution γ -ray radiography produced by a laser-plasma driven electron source. *Phys. Rev. Lett.* **94**, 025003 (2005).
11. Tsung, F. S. *et al.* Near-GeV-energy laser-wakefield acceleration of self-injected electrons in a centimeter-scale plasma channel. *Phys. Rev. Lett.* **93**, 185002 (2004).
12. Sun, G., Ott, E., Lee, Y. C. & Guzdar, P. Self-focusing of short intense pulses in plasmas. *Phys. Fluids* **30**, 526–532 (1987).
13. Faure, J. *et al.* Observation of laser pulse shortening in nonlinear plasma waves. *Phys. Rev. Lett.* **95**, 205003 (2005).
14. Hsieh, C.-T. *et al.* Tomography of injection and acceleration of monoenergetic electrons in a laser-wakefield accelerator. *Phys. Rev. Lett.* **96**, 095001 (2006).
15. Leemans, W. P. *et al.* GeV electron beams from a centimetre-scale accelerator. *Nature Phys.* **2**, 696–699 (2006).
16. Umstadter, D., Kim, J.-K. & Dodd, E. Laser injection of ultrashort electron pulses into wakefield plasma waves. *Phys. Rev. Lett.* **76**, 2073–2076 (1996).
17. Fubiani, G., Esarey, E., Schroeder, C. B. & Leemans, W. P. Beat wave injection of electrons into plasma waves using two interfering laser pulses. *Phys. Rev. E* **70**, 016402 (2004).
18. Kotaki, H., Masuda, S., Kando, M., Koga, J. K. & Nakajima, K. Head-on injection of a high quality electron beam by the interaction of two laser pulses. *Phys. Plasmas* **6**, 3296–3302 (2004).
19. Semuchin, S. & Malka, V. High density gas jet nozzle design for laser target production. *Rev. Sci. Instrum.* **72**, 2961–2965 (2001).
20. Glinec, Y. *et al.* Absolute calibration for a broadrange single shot electron spectrometer. *Rev. Sci. Instrum.* **77**, 103301 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank J.-P. Rousseau, F. Burgy, B. Mercier, A. Tazi, D. Douillet, T. Lefrou, F. Alahyane and J.-L. Charles. J.F. thanks S. Sebban for his patient help with the experiment. J.F. acknowledges discussions with G. Fubiani. We also acknowledge the support of the European Community Research Infrastructure Activity under the FP6 “Structuring the European Research Area” programme (CARE).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to V.M. (victor.malka@ensta.fr).

LETTERS

Thermal radiation scanning tunnelling microscopy

Yannick De Wilde^{1*}, Florian Formanek^{1*†}, Rémi Carminati², Boris Gralak³, Paul-Arthur Lemoine¹, Karl Joulain⁴, Jean-Philippe Mulet^{2†}, Yong Chen⁵ & Jean-Jacques Greffet²

In standard near-field scanning optical microscopy (NSOM), a subwavelength probe acts as an optical 'stethoscope' to map the near field produced at the sample surface by external illumination¹. This technique has been applied using visible^{1,2}, infrared³, terahertz⁴ and gigahertz^{5,6} radiation to illuminate the sample, providing a resolution well beyond the diffraction limit. NSOM is well suited to study surface waves such as surface plasmons⁷ or surface-phonon polaritons⁸. Using an aperture NSOM with visible laser illumination, a near-field interference pattern around a corral structure has been observed⁹, whose features were similar to the scanning tunnelling microscope image of the electronic waves in a quantum corral¹⁰. Here we describe an infrared NSOM that operates without any external illumination: it is a near-field analogue of a night-vision camera, making use of the thermal infrared evanescent fields emitted by the surface, and behaves as an optical scanning tunnelling microscope^{11,12}. We therefore term this instrument a 'thermal radiation scanning tunnelling microscope' (TRSTM). We show the first TRSTM images of thermally excited surface plasmons, and demonstrate spatial coherence effects in near-field thermal emission.

The TRSTM is based on a home-made apertureless infrared NSOM¹³ without external illumination. It is an atomic force microscope (AFM) with a hot sample holder combined with an infrared optical microscope and a mercury-cadmium-telluride (HgCdTe) detector. The AFM uses a tungsten tip, which oscillates perpendicularly to the sample surface at a frequency Ω . The tip periodically hits the sample surface, while its oscillation amplitude is stabilized by electronic feedback. Measurements of the feedback voltage, while scanning the sample under the tip, provide a topographic AFM image. Resistive heating of the sample holder gives the flexibility to raise the sample temperature up to 200 °C in order to increase thermal emission. The sample holder is thermally insulated with ceramics from the piezoelectric transducer used to scan the sample.

The tip acts as a scattering centre that radiates in the far field a signal linearly related to the infrared evanescent fields emitted by the surface^{8,14}. The near-field thermal emission scattered by the tip is collected by a Cassegrain objective with a numerical aperture NA = 0.5 mounted on a commercial optical microscope, and focused on a nitrogen-cooled infrared HgCdTe detector. Oscillations of the tip perpendicular to the sample surface result in a small oscillatory component on the detector signal, which is demodulated at frequency Ω or 2Ω using a lock-in amplifier. This TRSTM signal is measured as a function of tip position during the scans, producing a near-field thermal-emission image at the same time as the AFM image.

Most of the TRSTM measurements presented here have been performed on silicon carbide (SiC) samples with gold patterns. Both materials can support surface waves in the infrared—surface plasmon-polaritons (SPLPs) on gold, and surface phonon-polaritons (SPhPs) on SiC. SPhPs on SiC produce a resonance in the infrared, which has been investigated with an apertureless NSOM combined with tunable laser illumination⁸. Surface waves generate interesting coherence properties, and induce substantial field enhancement. Spatially coherent near-field thermal emission was predicted on SiC and gold¹⁵, and indirectly observed in far-field experiments^{16,17}. The existence of thermally excited surface waves also results in an enhancement of the electromagnetic energy density in the near field¹⁸ that has not been observed directly so far.

Gold patterns were prepared by optical lithography and evaporated on a flat SiC substrate to a thickness of 80 nm. Owing to the long decay length of surface waves on SiC (ref. 15), each pattern was separated from its nearest neighbour by 1.2 mm in order to avoid interactions. Figure 1 shows a photograph of a gold disk on SiC, along with a topographical (AFM) image and the corresponding TRSTM image, recorded simultaneously at a temperature $T = 170$ °C. The full spectral range (6.5–11.5 μm) of the HgCdTe detector was used, and the temperature was measured on the sample surface with a thermocouple. A contrast between gold and SiC is clearly observed in the TRSTM image, with a higher signal on gold. The boundary between the two materials is resolved within 100 nm. The resolution is two orders of magnitude better than what would be achieved using far-field infrared microscopy.

We have verified on the extremity of a gold stripe (Fig. 2a) that the TRSTM images are free of topographical artefacts¹⁹ (Fig. 2b). Nevertheless, when recording the TRSTM signal, another artefact may occur. As the tip is cooler than the sample, heat transfer occurs when the tip periodically hits the surface. This may produce a modulation of the sample temperature and therefore of its far-field thermal emission. It is thus relevant to check that a contrast in the TRSTM images survives when no heat transfer occurs between the tip and the sample—that is, when they are at the same temperature. Figure 2c shows an image recorded when the tip and the sample are both at room temperature. At this lower temperature, the TRSTM signal is reduced. Yet, the contrast between gold and SiC remains clearly visible. This proves that the origin of the signal is the conversion of thermally excited infrared evanescent fields into propagating waves by scattering. It might be surprising to observe a brighter image above gold than above SiC. Indeed, the local electromagnetic energy density in the near field of a flat air/SiC interface has a resonance at $\lambda = 10.6 \mu\text{m}$ (refs 8, 18), whereas the resonance for the air/gold

¹Laboratoire d'Optique Physique, Ecole Supérieure de Physique et de Chimie Industrielles, CNRS-UPR A0005, 10 rue Vauquelin, 75005 Paris, France. ²Laboratoire EM2C, Ecole Centrale Paris, CNRS, Grande Voie des Vignes, 92295 Châtenay-Malabry Cedex, France. ³Institut Fresnel, Faculté des Sciences et Techniques de St Jérôme, CNRS, case 161, Avenue Escadrille Normandie-Niemen, 13397 Marseille cedex 20, France. ⁴Laboratoire d'Etudes Thermiques, Ecole Nationale Supérieure de Mécanique et d'Aérotechnique, BP 109, 86960 Futuroscope Cedex, France. ⁵Laboratoire de Photonique et de Nanostructures, CNRS, Route de Nozay, 91460 Marcoussis, France, and Département de Chimie, Ecole Normale Supérieure, 24 Rue Lhomond, 75231 Paris Cedex 05, France. [†]Present addresses: Nanophotonics Laboratory-RIKEN, Wako, Saitama 351-0198, Japan (F.F.); Saint-Gobain Recherche, 39 Quai Lucien Lefranc, 93300 Aubervilliers, France (J.-P.M.).

*These authors contributed equally to this work.

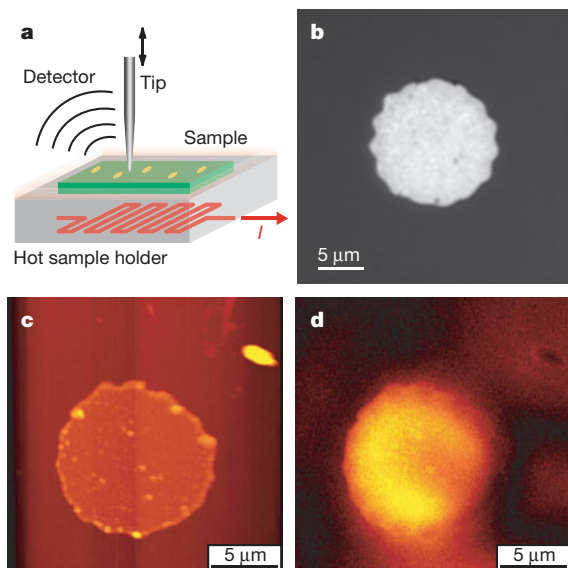


Figure 1 | Detection of the thermal radiation scanning tunnelling microscope (TRSTM) signal. Experimental sketch (a) and images (b, c, d) of gold disks evaporated on a SiC substrate. **b**, Image taken with a camera through a NA = 0.9 microscope objective with visible illumination. **c**, AFM topography of the gold structure on SiC, and **d**, TRSTM image recorded simultaneously at $T = 170^\circ\text{C}$, showing the infrared near-field thermal emission scattered by the tip. No external illumination was used for the recording of the TRSTM image. A bright fringe is observed along the rim of the disk.

interface takes place in the visible part of the spectrum. We now discuss the origin of the signal.

In order to understand the TRSTM images, we have studied a simple system consisting of a gold stripe with a variable width on a SiC substrate. Figure 3a shows results obtained at $T = 170^\circ\text{C}$ when placing an interference filter in front of the HgCdTe detector and demodulating the TRSTM signal at Ω . This filter performs band-pass filtering on the wavelength λ of the incoming radiation, centred at $\lambda = 10.9\ \mu\text{m}$ with a $1\ \mu\text{m}$ bandwidth. In the presence of the filter, fringes are observed on gold in the TRSTM images. Far from the extremities, the fringes preferentially align along the edges. This is a clear manifestation of interference of thermally excited near fields. It had been predicted theoretically that thermally excited surface waves induce long-range coherence¹⁵. This had been used to design a highly directional thermal source in the far field¹⁶. Here, we report the first (to our knowledge) direct experimental demonstration of spatial coherence of thermal emission in the near field. It is known

that gold supports SPLPs with a very long decay length in the infrared. We thus attribute the fringes to interferences of SPLPs on the gold stripe. The stripe behaves as a two-dimensional cavity for SPLPs. This explanation is supported by the fact that the number of fringes increases as the width of the stripe increases. To summarize, the system is a two-dimensional analogue of a Fabry–Perot cavity (the gold stripe) in a homogeneous medium (the SiC surface) in thermal equilibrium.

We now discuss the equivalence between TRSTM and scanning tunnelling microscopy (STM) and the implications for the interpretation of our measurements. In STM, electronic states in the sample are described by waves that decay exponentially in vacuum. The states are occupied according to Fermi–Dirac statistics. In the presence of a conducting tip, a tunnelling current is established and yields the signal. In TRSTM, electromagnetic surface states on the sample are described by waves that decay exponentially in vacuum. These states are occupied according to Bose–Einstein statistics. In the presence of a tip, light is scattered in the far field towards a detector. It has been shown that both processes can be described by the same formalism, assuming weak tip–sample coupling¹¹. It is known that within this assumption, an STM with a point-like tip measures the electronic local density of states²⁰ (eLDOS). As discussed in ref. 12, a point-like apertureless NSOM measuring thermally emitted fields probes the electromagnetic local density of states (EM-LDOS). Note that the STM yields the eLDOS at the Fermi energy, whereas TRSTM probes the EM-LDOS at a frequency that can be defined by a suitable filter. We stress that two conditions need to be fulfilled to measure the EM-LDOS. (1) All electromagnetic modes must be populated. (2) The detection process should have the same response for all modes. In our set-up, thermal excitation ensures that all available modes are populated according to Bose–Einstein statistics. However, the tip used in the set-up is not point-like. It selects chiefly one linear polarization, parallel to the tip axis^{12,14}. The corresponding projected EM-LDOS is the one that would determine the lifetime of an emitter with a transition dipole along the same axis²¹.

In order to analyse our data, we have performed a numerical calculation of the projected EM-LDOS above the gold stripe, at the resonance wavelength of SPhPs on SiC ($\lambda = 10.6\ \mu\text{m}$), using a procedure described in ref. 22. The calculated quantity is the projected EM-LDOS $\text{Im}(\mathbf{u} \cdot \mathbf{G} \cdot \mathbf{u})$, where $\mathbf{G}$ is the electric Green dyadic and $\mathbf{u}$ the unit vector along the direction of the tip axis, which is supposed to be perpendicular to the sample surface. For SPLPs on gold, the EM-LDOS is chiefly given by this quantity¹². The result is shown in Fig. 3. It is seen that close to the surface ($d = 200\ \text{nm}$), the EM-LDOS is dominated by the SPhPs on SiC. At a distance of $3\ \mu\text{m}$, the EM-LDOS is dominated by the SPLPs on gold whose decay length away from the surface is of the order of many wavelengths. We clearly observe a fringe structure, indicating that the stripe behaves as an SPLP cavity.

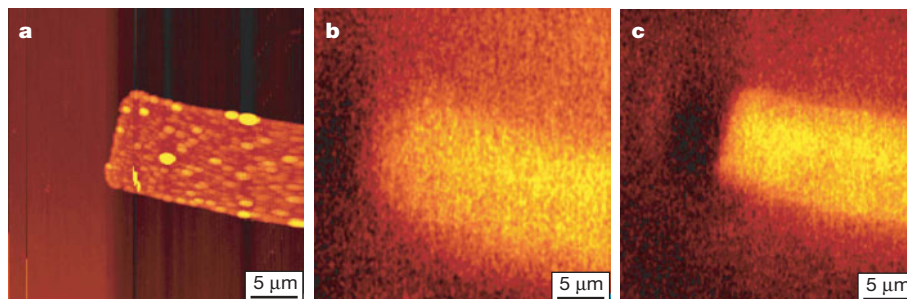


Figure 2 | Analysing the origin of the TRSTM signal. Images recorded at the extremity of a gold stripe on SiC when the tip and the sample are at room temperature. **a**, AFM topography. **b**, TRSTM image recorded with the tip scanning $5\ \mu\text{m}$ above the surface with the feedback loop disconnected. **c**, TRSTM image recorded while the tip periodically hits the sample surface with the feedback loop on ('tapping mode'). Despite the expected loss of resolution at a higher altitude due to the evanescent decay of the high spatial

frequencies of the field, the image in **b** is qualitatively the same as in **c** where the feedback loop is activated. This proves that the optical signal is not an optical read-out of the tip position. The possibility of imaging when the tip and the sample are in thermal equilibrium proves that the origin of the TRSTM signal is scattering of evanescent infrared fields confined close to the surface, and is not due to a local modulation of the sample temperature by the tip.

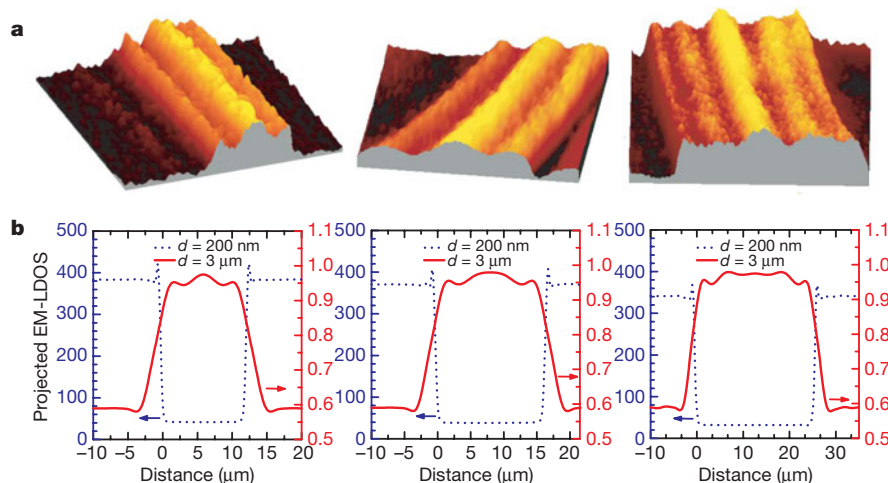


Figure 3 | EM-LDOS measurement by TRSTM. Images recorded at $T = 170^\circ\text{C}$ on three different areas of a 1-mm-long gold stripe on SiC, whose width varies continuously between the two extremities of the stripe from 10 to 30 μm . A band-pass filter (maximum at 10.9 μm ; width, 1 μm) performs energy selection (thus improving temporal coherence) on the radiation that reaches the infrared HgCdTe detector. **a**, Three-dimensional view of the TRSTM signal. From left to right, the width of the gold stripe is 12 μm , 16 μm and 25 μm . Fringes are observed on the gold stripe, giving a direct demonstration of the partial spatial coherence of the thermal radiation in the near field. Similar images (not shown) have been obtained on a gold stripe over a SiO_2 substrate, suggesting that the SPhPs on SiC do not play a key role.

A comparison of the calculation and the experimental data shows good agreement for the structure and the number of fringes above the gold stripe. These results clearly suggest that our instrument yields a signal that is related to the EM-LDOS averaged over the normal axis of the sample and that differs from the EM-LDOS close to the surface. Note that our numerical calculation also showed that there is no significant coupling between the EM-LDOS on SiC and the EM-LDOS on gold.

To understand these results, we have studied the z -dependence of the instrument. It has been shown that only a small part of the tip contributes to the signal^{8,23}. The effective size of this region depends on the tip geometry and on the frequency (Ω , 2Ω , ...) used for the signal demodulation^{8,23}. To get some insight about the vertical extension along which averaging of the EM-LDOS occurs, we have estimated the size of the region of the tip which contributes significantly to the TRSTM signal. To this end, we have measured the decay length of the signal d^* when the tip-sample distance is increased over a gold region. We typically find $d^* \approx 3 \mu\text{m}$ and 200 nm when demodulating the signal at Ω and 2Ω , respectively. Hence, demodulating the signal at 2Ω dramatically reduces the vertical extension from the sample surface along which the EM-LDOS is probed. Our images in Fig. 3 are consistent with the calculation of the EM-LDOS at 3 μm . We observe a qualitative agreement of the x - y dependence of the calculated EM-LDOS with the data, suggesting that the tip averages the signal over the normal axis. A point-like tip should be used to measure the z -dependence of the EM-LDOS. An image obtained when demodulating the signal at 2Ω is shown in Fig. 4a. It is seen that the signal is larger over SiC. This result is consistent with the EM-LDOS calculation at 200 nm, which shows a strong contribution of the SPhPs over SiC, and very low contrast of the fringes over gold, indicating that at very short distances, SPLPs do not yield the leading contribution.

To further check the role of SPhPs over SiC, we changed the wavelength from 10.9 μm to 8 μm so that the resonance frequency of SPhPs was no longer in the detection range. As expected, we observe in Fig. 4b that the signal over SiC decreases significantly. In order to check that the substrate does not play a key influence, we have also measured the EM-LDOS over a gold stripe deposited on a SiO_2 substrate using a filter centred at 10.9 μm (Fig. 4c). SiO_2 has no

b, Calculated EM-LDOS ($\lambda = 10.6 \mu\text{m}$) above the gold stripe at a height d of 200 nm and 3 μm ; values are scaled to 1.0 in vacuum. At $d = 200 \text{ nm}$, the EM-LDOS is dominated by the contribution of the large wavevector SPhPs on SiC. In contrast, at $d = 3 \mu\text{m}$, the EM-LDOS is dominated by the contribution of the SPLPs over gold whose decay length along the normal of the sample is much larger. At this frequency, there are many SPhP modes with large wavevectors yielding a field highly confined close to the interface. There are fewer SPLP modes on gold, but they have a much larger extension in vacuum above the interface. The contribution to the EM-LDOS is thus dominated by SPhPs on SiC at $d = 200 \text{ nm}$ and by SPLPs on gold at $d = 3 \mu\text{m}$.

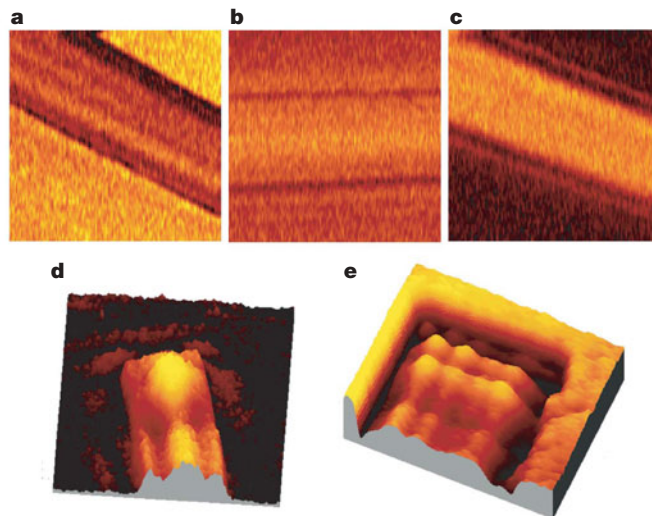


Figure 4 | EM-LDOS images. **a**, To check the role of SPhPs, we show an image taken of an 18- μm -wide gold stripe on a SiC substrate at frequency 2Ω with a band-pass filter at 10.9 μm (width 1 μm). At this frequency, we expect to detect a signal related to the EM-LDOS at roughly 200 nm above the surface. At this distance, the EM-LDOS is dominated by the resonant contribution of SPhPs over SiC (with a maximum at 10.6 μm). Indeed, we observe a signal larger over the SiC substrate than over the gold stripe. **b**, Image taken at 2Ω on a 25- μm -wide gold stripe on a SiC substrate but with a band-pass filter at 8 μm (width 1 μm). Within this detection range, there is no SPhP resonance. The signal over SiC is no longer dominant. **c**, Image taken under the same conditions as **a** but over a 21- μm -wide gold stripe on a SiO_2 substrate. SiO_2 has no SPhP resonance near 10.9 μm . It is seen that the signal is brighter on gold. In agreement with EM-LDOS calculations, no fringes are observed over gold because at this distance, the SPLP does not yield the leading contribution. **d**, EM-LDOS image taken at Ω at the edge of a gold stripe (10 μm width) on a SiC substrate. **e**, EM-LDOS image taken at Ω at the large end of the stripe (30 μm width). The figure shows the typical structure of the EM-LDOS in a cavity. The stripe extremity is a cavity for surface plasmons.

resonances at 10.9 μm , so that again the signal is larger over gold. When taking an image (not shown) demodulating at Ω , we retrieve the fringes over the gold stripe. This is an indication of a low influence of the substrate. Finally, we show in Fig. 4d and e the structure of the SPLPs close to the edges of the stripe. A more complicated structure due to the boundary conditions is revealed. It shows the EM-LDOS due to SPLPs in a two-dimensional cavity. Further work is needed to obtain a quantitative measurement of the EM-LDOS.

Received 21 February; accepted 19 September 2006.

- Pohl, D. W., Denk, W. & Lanz, M. Optical stethoscopy: Image recording with resolution $\lambda/20$. *Appl. Phys. Lett.* **44**, 651–653 (1984).
- Zenhausern, F., Martin, Y. & Wickramasinghe, H. K. Scanning apertureless microscopy: optical imaging at 10 Angstrom resolution. *Science* **269**, 1083–1085 (1995).
- Lahrech, A., Bachelot, R., Gleyzes, P. & Boccara, A. C. Infrared-reflection-mode near-field microscopy using an apertureless probe with a resolution of $\lambda/600$. *Opt. Lett.* **21**, 1315–1317 (1996).
- Chen, H.-T., Kersting, R. & Cho, G. C. Terahertz imaging with nanometer resolution. *Appl. Phys. Lett.* **83**, 309–311 (2003).
- Knoll, B., Keilmann, F., Kramer, A. & Guckenberger, R. Contrast of microwave near-field microscopy. *Appl. Phys. Lett.* **70**, 2667–2669 (1997).
- Ash, E. A. & Nicholls, G. Super-resolution aperture scanning microscope. *Nature* **237**, 510–512 (1972).
- Marti, O. *et al.* Near-field optical measurement of the surface plasmon field. *Opt. Commun.* **96**, 225–228 (1993).
- Hillenbrand, R., Taubner, T. & Keilmann, F. Phonon-enhanced light-matter interaction at the nanometre scale. *Nature* **418**, 159–161 (2002).
- Chicanne, C. *et al.* Imaging the local density of states of optical corrals. *Phys. Rev. Lett.* **88**, 097402 (2002).
- Crommie, M. F., Lutz, C. P. & Eigler, D. M. Confinement of electrons to quantum corrals on a metal surface. *Science* **262**, 218–220 (1993).
- Carminati, R. & Saenz, J. J. Scattering theory of Bardeen's formalism for tunneling: New approach to near-field microscopy. *Phys. Rev. Lett.* **84**, 5156–5159 (2000).
- Joulain, K., Carminati, R., Mulet, J.-P. & Greffet, J.-J. Definition and measurement of the local density of electromagnetic states close to an interface. *Phys. Rev. B* **68**, 245405 (2003).
- De Wilde, Y., Formanek, F. & Aigouy, L. Apertureless near-field scanning optical microscope based on a quartz tuning fork. *Rev. Sci. Instrum.* **74**, 3889–3891 (2003).
- Porto, J. A., Carminati, R. & Greffet, J.-J. Theory of electromagnetic field imaging and spectroscopy in scanning near-field optical microscopy. *J. Appl. Phys.* **88**, 4845–4850 (2000).
- Carminati, R. & Greffet, J.-J. Near-field effects in spatial coherence of thermal sources. *Phys. Rev. Lett.* **82**, 1660–1663 (1999).
- Greffet, J.-J. *et al.* Coherent emission of light by thermal sources. *Nature* **416**, 61–64 (2002).
- Kreiter, M. *et al.* Thermally induced emission of light from a metallic diffraction grating, mediated by surface plasmons. *Opt. Commun.* **168**, 117–122 (1999).
- Shchegrov, A. V., Joulain, K., Carminati, R. & Greffet, J.-J. Near-field spectral effects due to electromagnetic surface excitations. *Phys. Rev. Lett.* **85**, 1548–1551 (2000).
- Hecht, B., Bielefeldt, H., Inouye, Y., Pohl, D. W. & Novotny, L. Facts and artifacts in near-field optical microscopy. *J. Appl. Phys.* **81**, 2492–2498 (1997).
- Tersoff, J. & Hamann, D. R. Theory of the scanning tunneling microscope. *Phys. Rev. B* **31**, 805–813 (1985).
- Agarwal, G. S. Quantum electrodynamics in the presence of dielectrics and conductors. IV General theory for spontaneous emission in finite geometries. *Phys. Rev. A* **12**, 1475–1497 (1975).
- Gralak, B., de Dood, M. J. A., Tayeb, G., Enoch, S. & Maystre, D. Theoretical study of photonic band gaps in woodpile crystals. *Phys. Rev. E* **67**, 066601 (2003).
- Knoll, B. & Keilmann, F. Enhanced dielectric contrast in scattering-type scanning near-field optical microscopy. *Opt. Commun.* **182**, 321–328 (2000).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A.C. Boccara for comments and L. Aigouy for AFM measurements on the samples. This work was supported by the Ministère délégué à la Recherche (programme ACI Jeunes chercheurs) and the Centre National de la Recherche Scientifique.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to Y.D.W. (dewilde@optique.espci.fr).

LETTERS

Oxidation of the Ediacaran Ocean

D. A. Fike¹, J. P. Grotzinger^{1†}, L. M. Pratt² & R. E. Summons¹

Oxygenation of the Earth's surface is increasingly thought to have occurred in two steps. The first step, which occurred ~2,300 million years (Myr) ago, involved a significant increase in atmospheric oxygen concentrations and oxygenation of the surface ocean^{1,2}. A further increase in atmospheric oxygen appears to have taken place during the late Neoproterozoic period^{3,4} (~800–542 Myr ago). This increase may have stimulated the evolution of macroscopic multicellular animals and the subsequent radiation of calcified invertebrates^{4,5}, and may have led to oxygenation of the deep ocean⁶. However, the nature and timing of Neoproterozoic oxidation remain uncertain. Here we present high-resolution carbon isotope and sulphur isotope records from the Huqf Supergroup, Sultanate of Oman, that cover most of the Ediacaran period (~635 to ~548 Myr ago). These records indicate that the ocean became increasingly oxygenated after the end of the Marinoan glaciation, and they allow us to identify three distinct stages of oxidation. When considered in the context of other records from this period^{7–15}, our data indicate that certain groups of eukaryotic organisms appeared and diversified during the second and third stages of oxygenation. The second stage corresponds with the Shuram excursion in the carbon isotope record¹⁶ and seems to have involved the oxidation of a large reservoir of organic carbon suspended in the deep ocean⁶, indicating that this event may have had a key role in the evolution of eukaryotic organisms. Our data thus provide new insights into the oxygenation of the Ediacaran ocean and the stepwise restructuring of the carbon^{6,16,17} and sulphur cycles^{3,18,19} that occurred during this significant period of Earth's history.

The Huqf Supergroup provides one of the best-preserved, most continuous sections of Ediacaran-age strata²⁰ (Fig. 1). The Abu Mahara Group contains Marinoan-equivalent (~635-Myr-old) glacial deposits (Fiq Formation and associated Hadash cap carbonate) that overlie ~800-Myr-old crystalline basement⁷. Nafun Group sediments were deposited in a regionally extensive sag basin under open, shallow marine conditions, and each formation can be traced laterally for several hundred kilometres²¹. Nafun strata (see Supplementary Information for a detailed discussion of lithostratigraphy) comprise two clastic-to-carbonate shallowing-upward successions (Masirah Bay Formation–Khufai Formation; Shuram Formation–Buah Formation) with an unconformity across the Khufai–Shuram boundary that probably includes the interval of Gaskiers glaciation ~580 Myr ago (ref. 22). Global correlation of $\delta^{13}\text{C}_{\text{carb}}$ (carbonate, see Supplementary Information for carbon and sulphur isotope nomenclature) anomalies provides two age constraints for the Buah Formation (Fig. 1): ~550 Myr for the mid-Buah (correlation with Doushantuo Formation, China^{7,9}), and ~548 Myr for the upper Buah (correlation with Nama Group, Namibia^{7,8}). These correlations are supported by multiple ages⁷ spanning 541–547 Myr ago obtained from the overlying Ara Group.

Here we present carbon and sulphur isotope data from the Huqf Supergroup (Fig. 1; see Supplementary Information for detailed discussion of the data). All data were collected from cuttings in the Miqrat-1 well (Supplementary Fig. S1), previously established as one of the most representative subsurface sections of Huqf strata¹⁶. The Shuram Formation preserves a >15‰ negative excursion in $\delta^{13}\text{C}_{\text{carb}}$, reaching a minimum of about –12‰. The excursion spans ~500 m of stratigraphic section¹⁶ and is the largest known perturbation to the global carbon cycle in Earth history.

The Shuram excursion is fundamentally different from all known $\delta^{13}\text{C}_{\text{carb}}$ excursions (for example the Marinoan cap carbonate, the Ediacaran/Cambrian boundary and the Permian/Triassic boundary), in that $\delta^{13}\text{C}_{\text{carb}}$ reaches well below the mantle value of about –6‰. Because $\delta^{13}\text{C}_{\text{carb}}$ values below –6‰ cannot readily be explained by changes in organic carbon burial or isotope fractionation during carbon fixation⁴, the Shuram excursion was initially attributed to diagenetic alteration of the $\delta^{13}\text{C}_{\text{carb}}$ signal^{16,21}. However, standard methods of assessing diagenesis (see Supplementary Fig. S2a, b, and discussion in Supplementary Information) indicate that these samples retain primary $\delta^{13}\text{C}_{\text{carb}}$ values. Furthermore, the Shuram excursion is now documented in more than 30 sections in Oman, covering a region in excess of 10⁵ km² with little variation in either magnitude or duration^{16,21,23}. Additionally, potentially correlative excursions have been found across the globe: the Doushantuo Formation, China⁹; the Wonoka Formation, Australia¹³; the Johnnie Formation, USA¹⁴; and the Nama Group, Namibia¹⁵ (Fig. 2). Taken together, they indicate strongly that the Shuram excursion is a primary record of an unprecedented perturbation to the global carbon cycle.

Unlike younger carbon isotopic anomalies that are expressed by $\delta^{13}\text{C}_{\text{carb}}$ in carbonate sections and $\delta^{13}\text{C}_{\text{org}}$ (organic carbon) in siliciclastic sections, the Shuram excursion has been identified only in carbonate sections^{9,13–16,24}. Given the absence of covariation in $\delta^{13}\text{C}_{\text{org}}$ with $\delta^{13}\text{C}_{\text{carb}}$ during the Shuram excursion reported here and from the Wonoka Formation in Australia¹³, the excursion would be undetectable in Shuram-equivalent siliciclastic strata, which require $\delta^{13}\text{C}_{\text{org}}$ for chemostratigraphy. This may explain why the Shuram excursion has not been found in more Ediacaran sections (for example siliciclastic-dominated sections such as the Windermere Supergroup, Canada).

The data presented here indicate progressive oxidation of the Ediacaran ocean in three stages after the end of Marinoan glaciation. These successive stages occur well after the last of the extreme Neoproterozoic glaciations⁹ and before the extinction and subsequent evolutionary radiation across the Ediacaran/Cambrian boundary²⁰. The first stage of oxidation occurs in the Khufai and Masirah Bay formations above the Marinoan-equivalent Hadash cap carbonate. In the Hadash cap, $\Delta\delta^{34}\text{S}$, the difference between coeval $\delta^{34}\text{S}_{\text{SO}_4}$ (carbonate-associated sulphate) and $\delta^{34}\text{S}_{\text{pyr}}$ (pyrite)

¹Department of Earth, Atmospheric, and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA. ²Department of Geological Sciences, Indiana University, Bloomington, Indiana 47405, USA. [†]Present address: Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA.

(Fig. 1e), ranges from 1‰ to 12‰. These low values, consistent with those found above the coeval Marinoan diamictite in Namibia¹⁹, indicate marine sulphate concentrations ($[SO_4]$) of less than 200 μM (ref. 25), which prevents the expression of significant isotopic fractionation during bacterial sulphate reduction (BSR). Throughout the Masirah Bay and Khufai formations above the cap carbonate, $\Delta\delta^{34}S$ increases gradually to ~ 35 ‰, recording an increase to $[SO_4] > 200 \mu M$ and unlimited expression of BSR fractionation. Although it is difficult to gauge the magnitude of the increase in $[SO_4]$, reports^{18,19} of Ediacaran and early Cambrian enrichment and variability in $\delta^{34}S_{SO_4}$ suggest that $[SO_4]$ did not exceed ~ 5 mM (ref. 26), which is appreciably lower than the modern value of 28 mM. Two key fluxes for regulating $[SO_4]$ are riverine sulphate derived from pyrite oxidation (source) and marine reduction of sulphate to sulphide (sink). Because these fluxes both depend directly on oxygen concentrations, increased $[SO_4]$ probably correlates with increased oxygen availability. It is likely that the increase in $[SO_4]$ above the cap carbonate is in part the recovery from a glacially induced drawdown of sulphate concentrations¹⁹. However, on the basis of ~ 610 -Myr-old detrital zircons in the upper Khufai^{7,27} the minimum time represented by the increase in $\Delta\delta^{34}S$ in stage I strata overlying the ~ 635 -Myr-old Marinoan cap carbonate is ~ 25 Myr and therefore probably records an increase in atmospheric oxygen in addition to any effects associated with deglaciation. This increase in $\Delta\delta^{34}S$ constitutes the first stage (Fig. 1, stage I) of Ediacaran oxidation observed in the Huqf sediments. Throughout stage I strata, $\delta^{13}C_{carb}$ and $\delta^{13}C_{org}$ were not coupled.

The second stage (Fig. 1, stage II) of Ediacaran oxidation corresponds to the duration of the Shuram excursion and encompasses the Shuram and lower Buah formations. Although these strata lie immediately above the Khufai, the likely presence of a significant

unconformity prevents meaningful comparison between data from the base of the Shuram and those from the top of the Khufai. The absence of a parallel negative excursion in $\delta^{13}C_{org}$ suggests that pools of organic carbon and dissolved inorganic carbon in the Ediacaran ocean were effectively decoupled, such that excursions in $\delta^{13}C_{carb}$, which reflects the dissolved inorganic carbon pool—the source for carbon fixation of coeval organic matter—were not recorded in $\delta^{13}C_{org}$ of syndepositional sediments⁶. Our trends agree with the observation⁶ of an overall absence of covariation between $\delta^{13}C_{org}$ and $\delta^{13}C_{carb}$ in a global compilation of Neoproterozoic data from ~ 730 to 555 Myr ago. With decoupled carbon reservoirs, negative excursions in $\delta^{13}C_{carb}$ can result from the oxidation of part of a much larger DOC reservoir represented by the deep ocean, and their amplitude is limited by $\delta^{13}C_{org}$ (about -30 ‰) rather than mantle composition (about -6 ‰) as in a steady-state model. (Here and throughout the text, the term DOC is used to refer to a large reservoir of organic carbon (not the result of coeval primary production) suspended in the deep ocean; use of the term DOC is not intended to distinguish between truly ‘dissolved’ organic carbon and suspended colloidal or fine particulate organic carbon.) This vast organic reservoir would effectively buffer coexisting $\delta^{13}C_{org}$ through adsorption on siliciclastic particles and/or during carbonate precipitation in significant excess of the influx of detritus from primary production. At the initiation of the Shuram excursion, this adsorption would have to contribute $\geq 90\%$ of TOC to mask the signal of $\delta^{13}C$ -depleted primary organic matter. However, because the organic reservoir is depleted through progressive oxidation, the relative contribution of coeval $\delta^{13}C$ -depleted primary production to $\delta^{13}C_{org}$ increases; this is observed (Fig. 1b) in the upper portion of stage II as $\delta^{13}C_{org}$ decreases from about -26 ‰ to about -31 ‰. As the organic reservoir is finally oxidized, the onset of covariation in

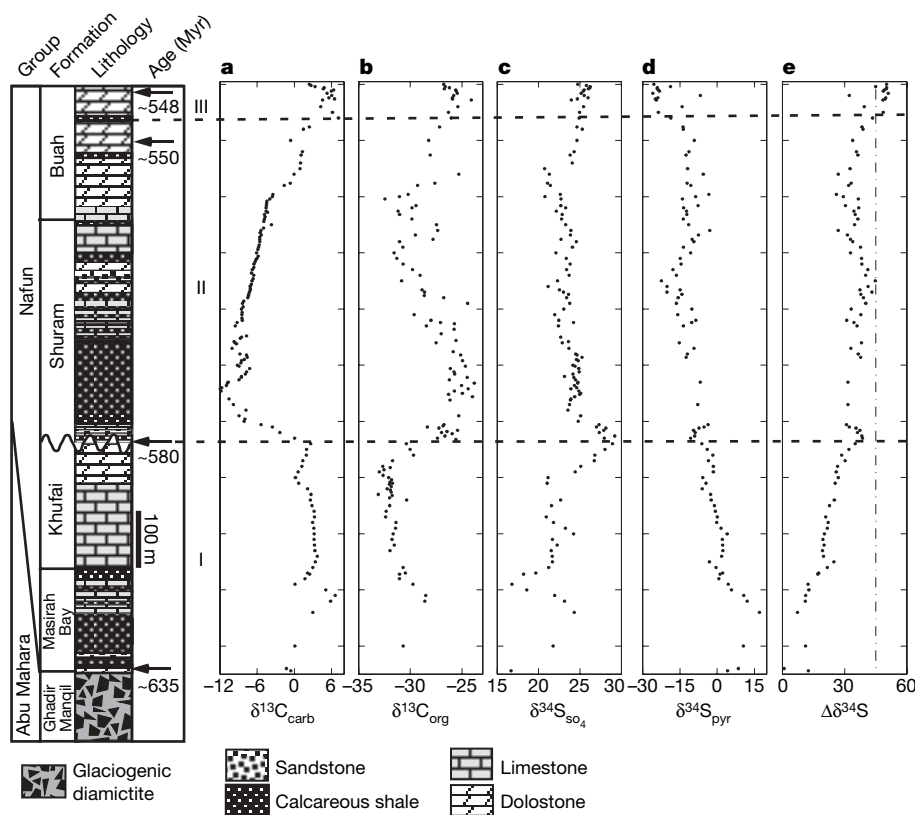


Figure 1 | Huqf Supergroup showing geochronological, palaeobiological and chemostratigraphic constraints. The Hadash cap carbonate lies immediately above the Ghadir Manqil glacial diamictite and conformably below the Masirah Bay Formation. Geochronological constraints are from correlation to other sections containing U–Pb zircon ages on ash beds^{8,9,22}.

Data plotted are $\delta^{13}C_{carb}$ (a), $\delta^{13}C_{org}$ (TOC) (b), $\delta^{34}S_{SO_4}$ (c), $\delta^{34}S_{pyr}$ (d) and $\Delta\delta^{34}S = \delta^{34}S_{SO_4} - \delta^{34}S_{pyr}$ (e). The dashed line at 46‰ indicates the maximum fractionation associated with BSR. The 2σ error bars for isotopic measurements, based on replicate analyses of standards and samples, are $\delta^{13}C_{carb}$ and $\delta^{34}S_{SO_4} \leq 0.15$ ‰; $\delta^{13}C_{org}$ and $\delta^{34}S_{pyr} \leq 0.4$ ‰.

$\delta^{13}\text{C}_{\text{carb}}$ and $\delta^{13}\text{C}_{\text{org}}$ ($r^2 = 0.72$ for the upper ~ 200 m of section) becomes apparent. Further evidence supporting oxidation during stage II is given in Supplementary Information.

The constancy of $\Delta\delta^{34}\text{S}$ during the Shuram excursion indicates that, despite these changes, BSR under sulphate-replete conditions remained the dominant pathway for sulphur cycling in the ocean. However, in the upper Buah, an increase in $\Delta\delta^{34}\text{S}$ to $\sim 50\text{‰}$ indicates a change in the metabolisms involved in marine sulphur cycling. The maximum fractionation observed¹³ in both laboratory and field studies during BSR is 46‰; typical fractionations are often much smaller²⁸. However, communities characterized by BSR coupled with bacterial sulphur disproportionation (BSD), a pathway³ in which intermediate-valence sulphur species (for example S^0 and $\text{S}_2\text{O}_3^{2-}$) are split into ^{34}S -enriched sulphate and ^{34}S -depleted hydrogen sulphide, have fractionations that can approach 70‰. There is evidence from paired $\delta^{33}\text{S}$ – $\delta^{34}\text{S}$ data for the evolution of BSD as early as the Mesoproterozoic²⁹; however, there have been no reports of a coeval $\Delta\delta^{34}\text{S}$ of more than 46‰ in any Proterozoic sample³⁰. Because BSD requires the presence of intermediate-valence sulphur species, the oxidative sulphur cycle must have been active. Evidence for BSD therefore demonstrates additional oxidation of the Ediacaran ocean immediately after the Shuram recovery and constitutes the third stage (Fig. 1, stage III) of Ediacaran oxidation preserved in the Huqf sediments.

The data presented in this paper provide a record of multistage oxidation of Earth's surficial environment during Ediacaran time. We now place this record into a global palaeobiological context by integrating coeval records of both plankton (acanthomorph acritarchs) and benthic Ediacaran soft-bodied animals into our chemostratigraphic

context (Fig. 2). Well-characterized palynological assemblages that record the evolution of the acanthomorph acritarchs¹⁰ are preserved in two chemostratigraphically constrained sections (China and Australia) potentially correlative to the Shuram excursion (Fig. 2). In Australia, the appearance of acanthomorph acritarchs coincides approximately with the onset of the Shuram excursion (our stage II oxidation) in the Wonoka Formation, Adelaide Rift Complex^{10,13}. The Acraman impact event has previously been postulated as the stimulus for the acanthomorph acritarch radiation¹⁰. However, impact ejecta occur ~ 140 m below strata marking both the radiation and the initiation of the Shuram excursion. We therefore propose alternatively that the acritarch radiation is temporally correlated with the increased oxidation of the ocean and subsequent ecological changes that occurred during the Shuram excursion. The most diverse assemblage of acritarchs occurs in the Julie Formation, Amadeus Basin¹⁰, during an excursion in $\delta^{13}\text{C}_{\text{carb}}$ of up to +5‰ (Fig. 2), which we interpret as equivalent to the stage III oxidation event based on $\delta^{13}\text{C}_{\text{carb}}$ correlation to the upper Buah Formation in Oman and the lower Nama Group of Namibia, where the age of the correlative positive $\delta^{13}\text{C}_{\text{carb}}$ isotopic excursion is constrained to ~ 548 Myr (ref. 8). Although neither the carbonates of the Nama Group nor those of the Buah Formation have yielded well-preserved acritarch populations, acanthomorph acritarchs in the shale-rich Doushantuo Formation of China¹⁰ appear at about the same stratigraphic position as the reported Shuram excursion⁹. The coincidence of the latter two stages of oxidation with the first appearance and subsequent diversification of acanthomorph acritarchs suggests that the oxidation of the ocean had a major role in the evolution of these planktic photosynthetic organisms.

It is not possible to determine confidently whether the first known appearance of the Ediacaran biota at ~ 575 Myr ago (ref. 12) predates or postdates the onset of the Shuram excursion, because of a lack of overlapping biostratigraphic and chemostratigraphic data. However, both the Shuram excursion and the first appearance of Ediacaran organisms are believed to postdate the Gaskiers glacial event ~ 580 Myr ago (see Supplementary Information), yet predate the positive $\delta^{13}\text{C}_{\text{carb}}$ excursion ~ 548 Myr ago. It is therefore possible that the second stage of oxygenation helped to stimulate the radiation of the first Ediacaran organisms, which are interpreted to have been sessile frond-like animals (such as *Charnia*); although these organisms seem to have inhabited a variety of different palaeoenvironments^{11,12}, all evidence indicates that they were probably confined to the oxygenated mixed layer of the ocean. However, it is not until ~ 555 Myr ago (ref. 11), as the Shuram excursion drew to a close, that we see the appearance of the first motile, bilaterian organisms (for example, *Kimberella*) in a wider range of depositional facies. Finally, the first calcifying metazoa *Cloudina* and *Namacalathus*, well documented in both Oman and Namibia, appeared ~ 548 Myr ago, after the strong BSD signal. The coincidence of the latter two stages of oxidation identified in this study with the appearance of motile and calcifying Ediacarans, respectively, suggests that the events observed in Oman were global and affected the evolution of early metazoa.

Examination of the Huqf Supergroup, Sultanate of Oman, reveals a three-stage oxidation of the Ediacaran ocean. The first stage corresponds to an increase in sulphate concentrations ($>200 \mu\text{M}$) after the Marinoan glaciation, due in part to an increase in atmospheric oxygen. The second stage consists of the Shuram excursion and the oxidation of a deep-ocean DOC reservoir, probably the last major redox barrier to ocean oxygenation. This stage is coincident with the appearance of complex acanthomorph acritarchs and motile metazoa. The final stage of Ediacaran oxidation is characterized by strong signals of BSD and covariation in $\delta^{13}\text{C}_{\text{carb}}$ and $\delta^{13}\text{C}_{\text{org}}$, coincident with an increase in the diversity of acanthomorph acritarchs and the first appearance of the calcifying metazoa *Namacalathus* and *Cloudina*. Taken together, these data record the progressive oxygenation of the Ediacaran ocean, stimulating the evolution of both planktic and benthic groups of organisms.

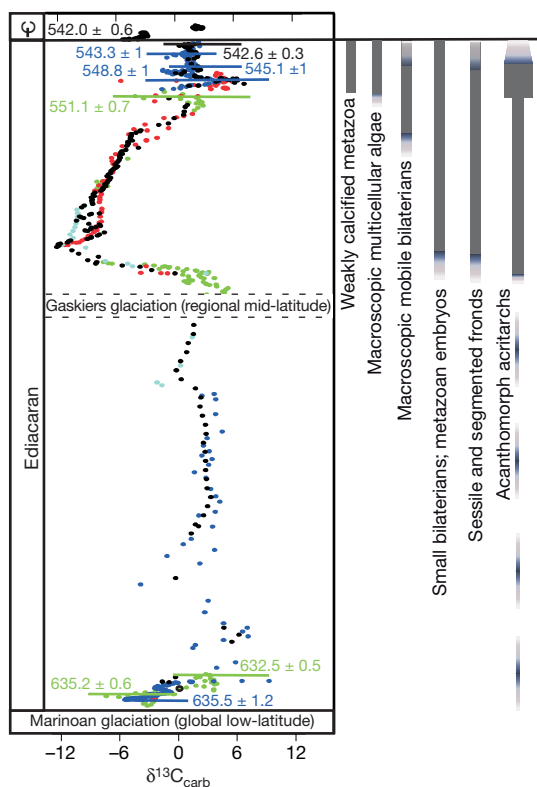


Figure 2 | Compilation of Ediacaran $\delta^{13}\text{C}_{\text{carb}}$ from sections known to contain the Shuram excursion. The vertical axis is approximate time, with known geochronological constraints plotted. On the right are plotted the known range of acanthomorph acritarchs and Ediacaran biota (shading indicates uncertainty relative to $\delta^{13}\text{C}$ chemostratigraphy). Points are coloured as follows: black, Oman; dark blue, Namibia; green, China; red, Australia; cyan, Death Valley. Full references for data, U/Pb zircon ages, biostratigraphic range, and the method for correlating sections are provided in Supplementary Information.

Received 29 November 2005; accepted 12 October 2006.

- Holland, H. D. *The Chemical Evolution of the Atmosphere and Oceans* (Princeton Univ. Press, Princeton, NJ, 1984).
- Bekker, A. *et al.* Dating the rise of atmospheric oxygen. *Nature* **427**, 117–120 (2004).
- Canfield, D. E. & Teske, A. Late Proterozoic rise in atmospheric oxygen concentration inferred from phylogenetic and sulphur-isotope studies. *Nature* **382**, 127–132 (1996).
- Des Marais, D. J., Strauss, H., Summons, R. E. & Hayes, J. M. Carbon isotope evidence for the stepwise oxidation of the Proterozoic environment. *Nature* **359**, 605–609 (1992).
- Knoll, A. H. & Carroll, S. B. Early animal evolution; emerging views from comparative biology and geology. *Science* **284**, 2129–2137 (1999).
- Rothman, D. H., Hayes, J. M. & Summons, R. E. Dynamics of the Neoproterozoic carbon cycle. *Proc. Natl Acad. Sci. USA* **100**, 8124–8129 (2003).
- Bowring, S. A., Grotzinger, J. P., Condon, D. J., Ramezani, J. & Newall, M. Geochronologic constraints on the chronostratigraphic framework of the Neoproterozoic Huqf Supergroup, Sultanate of Oman. *Am. J. Sci.* (in the press).
- Grotzinger, J. P., Bowring, S. A., Saylor, B. Z. & Kaufman, A. J. Biostratigraphic and geochronological constraints on early animal evolution. *Science* **270**, 598–604 (1995).
- Condon, D. *et al.* U–Pb Ages from the Neoproterozoic Doushantuo Formation, China. *Science* **308**, 95–98 (2005).
- Grey, K. *Ediacaran Palynology of Australia* Vol. 31 (Association of Australasian Paleontologists, Canberra, 2005).
- Martin, M. W. *et al.* Age of Neoproterozoic bilaterian body and trace fossils, White Sea, Russia: implications for metazoan evolution. *Science* **288**, 841–845 (2000).
- Narbonne, G. M. The Ediacara Biota: Neoproterozoic origin of animals and their ecosystems. *Annu. Rev. Earth Planet. Sci.* **33**, 1–22 (2005).
- Calver, C. R. Isotope stratigraphy of the Ediacarian (Neoproterozoic III) of the Adelaide Rift Complex, Australia, and the overprint of water column stratification. *Precamb. Res.* **100**, 121–150 (2000).
- Corsetti, F. A. & Kaufman, A. J. Stratigraphic investigations of carbon isotope anomalies and Neoproterozoic ice ages in Death Valley, California. *Geol. Soc. Am. Bull.* **115**, 916–932 (2003).
- Workman, R. K., Grotzinger, J. P. & Hart, S. R. Constraints on Neoproterozoic ocean chemistry from $\delta^{13}\text{C}$ and $\delta^{11}\text{B}$ analyses of carbonates from the Witvlei and Nama Groups, Namibia. *Geochim. Cosmochim. Acta* **66**, 847 (2002).
- Burns, S. J. & Matter, A. Carbon isotopic record of the latest Proterozoic from Oman. *Ecol. Geol. Helv.* **86**, 595–607 (1993).
- Logan, G. A., Hayes, J. M., Hieshima, G. B. & Summons, R. E. Terminal proterozoic reorganization of biogeochemical cycles. *Nature* **376**, 53–56 (1995).
- Canfield, D. E. The evolution of the Earth surface sulfur reservoir. *Am. J. Sci.* **304**, 839–861 (2004).
- Hurtgen, M. T. Sulfur cycling in the aftermath of a Neoproterozoic (Marinoan) snowball glaciation: Evidence for a syn-glacial sulfidic deep ocean. *Earth Planet. Sci. Lett.* **245**, 551–570 (2006).
- Amthor, J. E. *et al.* Extinction of *Cloudina* and *Namacalathus* at the Precambrian–Cambrian boundary in Oman. *Geology* **31**, 431–434 (2003).
- McCarron, G. *The Sedimentology and Chemostratigraphy of the Nafun Group, Huqf Supergroup, Oman*. Thesis, Univ. Oxford (2000).
- Bowring, S. A., Myrow, P. M., Landing, E. & Ramezani, J. Geochronological constraints on terminal Neoproterozoic events and the rise of Metazoans. *Astrobiology* **2**, 112 (2002).
- Le Guerroue, E., Allen, P. A. & Cozzi, A. Chemostratigraphic and sedimentological framework of the largest negative carbon isotopic excursion in Earth history: The Neoproterozoic Shuram Formation (Nafun Group, Oman). *Precamb. Res.* **146**, 68–92 (2006).
- Melezhik, V. A., Fallick, A. E. & Pokrovsky, B. G. Enigmatic nature of thick sedimentary carbonates depleted in ^{13}C beyond the canonical mantle value: The challenges to our understanding of the terrestrial carbon cycle. *Precamb. Res.* **137**, 131–165 (2005).
- Habicht, K. S., Gade, M., Thamdrup, B., Berg, P. & Canfield, D. E. Calibration of sulfate levels in the Archean Ocean. *Science* **298**, 2372–2374 (2002).
- Kah, L. C., Lyons, T. W. & Frank, T. D. Low marine sulphate and protracted oxygenation of the Proterozoic biosphere. *Nature* **431**, 834–838 (2004).
- Le Guerroue, E., Allen, P. A., Cozzi, A., Etienne, J. L. & Fanning, M. 50 million year duration negative carbon isotope excursion in the Ediacaran ocean. *Terra Nova* **18**, 147–153 (2006).
- Detmers, J., Bruchert, V., Habicht, K. S. & Kuever, J. Diversity of sulfur isotope fractionations by sulfate-reducing prokaryotes. *Appl. Environ. Microbiol.* **67**, 888–894 (2001).
- Johnston, D. T. *et al.* Active microbial sulfur disproportionation in the Mesoproterozoic. *Science* **310**, 1477–1479 (2005).
- Hurtgen, M. T., Arthur, M. A. & Halverson, G. P. Neoproterozoic sulfur isotopes, the evolution of microbial sulfur species, and the burial efficiency of sulfide as sedimentary pyrite. *Geology* **33**, 41–44 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Canfield for use of laboratory facilities and discussions; C. Colanero, J. Fong and S. Studley for laboratory assistance; A. Bradley, D. Finkelstein, G. Love, B. McElroy, A. Maloof and W. Watters for comments; and T. Lyons for suggestions that improved the manuscript. We thank Petroleum Development Oman (PDO) for access to samples and support for this project, and the Oman Ministry of Oil and Gas for permission to publish this paper. Support was provided by PDO and the National Aeronautics and Space Administration. J.P.G. and D.A.F. were supported by the Agouron Institute. L.M.P. was supported by a NASA Astrobiology Institute grant. R.E.S. was supported by an NSF Biocomplexity grant and a NASA Exobiology grant.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.A.F. (dfike@mit.edu).

LETTERS

How Neanderthal molar teeth grew

Roberto Macchiarelli¹, Luca Bondioli², André Debénath³, Arnaud Mazurier^{1,4}, Jean-François Tournepiche^{5,6}, Wendy Birch⁷ & Christopher Dean⁷

Growth and development are both fundamental components of demographic structure and life history strategy. Together with information about developmental timing they ultimately contribute to a better understanding of Neanderthal extinction. Primate molar tooth development tracks the pace of life history evolution most closely^{1,2}, and tooth histology reveals a record of birth as well as the timing of crown and root growth. High-resolution micro-computed tomography now allows us to image complex structures and uncover subtle differences in adult tooth morphology that are determined early in embryonic development³. Here we show that the timing of molar crown and root completion in Neanderthals matches those known for modern humans but that a more complex enamel–dentine junction morphology and a late peak in root extension rate sets them apart. Previous predictions about Neanderthal growth, based only on anterior tooth surfaces^{4,5}, were necessarily speculative. These data are the first on internal molar microstructure; they firmly place key Neanderthal life history variables within those known for modern humans.

To provide information about the morphology of the Neanderthal enamel–dentine junction (EDJ) and about the timing of crown and root formation, we selected two minimally worn molars from the site

of La Chaise-de-Vouthon, Charente, France⁶: a lower right deciduous second molar (S14-5) from the Riss III level (OIS 6) of Abri Suard, and a lower left permanent first molar (BD-J4-C9) from the Riss-Würm level (OIS 5e) of Abri Bourgeois-Delaunay (Supplementary Information).

A record of growth exists in the enamel and dentine that allows us to reconstruct their developmental history and the timing of crown and root formation^{7,8}. Some morphological features of the crown are the result of superficial overgrowths at the enamel surface toward the end of formation as a result of a longer secretory lifespan of ameloblasts. Others have an earlier developmental history and are mapped out at the EDJ in the embryonic tooth germ³. These are formed under the control of a signalling centre, the enamel knot, which directs differential proliferation and folding along the inner enamel epithelium^{9,10}.

We first used high-resolution synchrotron radiation micro-computed tomography (SR- μ CT) images to reveal the EDJ within both Neanderthal molars (Fig. 1). The Neanderthal metaconid or mid-trigonid crest, a typically diagnostic feature of Neanderthal molars¹¹, emerges as a distinct structure that is initially defined during early tooth germ development. This seems to be associated with a generally

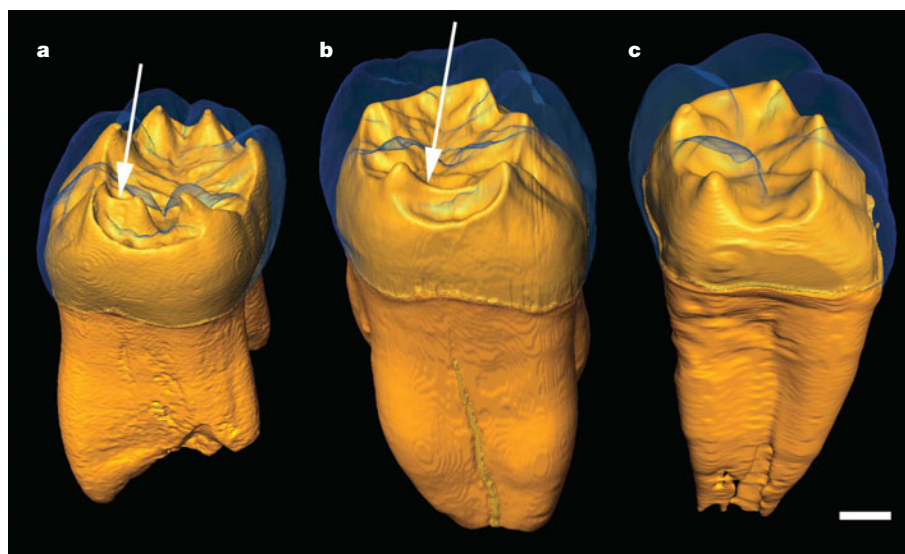


Figure 1 | SR- μ CT-based three-dimensional virtual reconstruction of deciduous and permanent Neanderthal molars from La Chaise. The transparent enamel caps are compared with a modern human permanent molar (slightly oblique mesiobuccal views). **a**, Neanderthal lower right deciduous second molar (LRm2). **b**, Neanderthal lower left permanent first molar (LLM1). **c**, Modern human lower M1. In each of the molars from La

Chaise, the metaconid, or mid-trigonid, crest (arrow), typical of Neanderthals but rare in modern humans, runs in both dentine and enamel between the two anterior cusps (metaconid and protoconid); in the deciduous molar the crest is divided¹¹. The feature is absent from the modern human specimen. Scale bar, 2 mm.

¹Laboratoire de Géobiologie, Biochronologie et Paléontologie Humaine, UMR 6046 CNRS, Université de Poitiers, 86022 Poitiers, France. ²Sezione di Antropologia, Museo Nazionale Preistorico Etnografico 'L. Pigorini', 00144 Rome, Italy. ³Université de Perpignan, 66000 Perpignan, France. ⁴Etudes Recherches Matériaux, 86022 Poitiers, France. ⁵Musée d'Angoulême, 16000 Angoulême, France. ⁶UMR 5199 CNRS, Université Bordeaux 1, 33405 Bordeaux, France. ⁷Department of Anatomy and Developmental Biology, University College London, London WC1E 6BT, UK.

more complex EDJ topography relative to the enamel surface area in both deciduous and permanent Neanderthal teeth (Supplementary Table 1). The percentage ratio between EDJ area and enamel surface area suggests that both modern human and Neanderthal deciduous molars have proportionately greater EDJ surface areas than permanent molars. However, both deciduous and permanent Neanderthal molars show a more complex EDJ than is typical of modern humans¹², with about 10% greater surface area. Larger sample sizes will be required to confirm this and focus further on how this might relate to function^{13,14}. We then used thin ground sections of the same molar teeth and polarized light microscopy to compare the subsequent timing of enamel formation, crown completion and root formation with times known for modern humans of diverse geographical origins^{15,16}.

There has been speculation, but no data, for the timing of molar tooth formation and gingival emergence in Neanderthals^{17–20}. We made direct counts and measurements of daily enamel cross-striations to retrieve information about the timing and rate of enamel formation. Birth is marked in deciduous teeth and first permanent molars by an accentuated neonatal line. The position of this line depends upon the time of initial mineralization *in utero* as well as on the time of birth. Enamel secretion rates through the first-formed cuspal enamel of the Neanderthal deciduous molar (Fig. 2) show a slight gradient from the EDJ to the enamel surface (increasing as modern humans do from between 2 or 3 μm per day to about 4 μm per day) but with the typical postnatal hypoplasia and decrease in daily rates of formation during the immediate period after birth. The position of the neonatal line and duration of this reaction to the changing physiology at birth is similar to that in modern humans²¹.

Enamel secretion rates through the first-formed cuspal regions of the Neanderthal permanent molar teeth (Fig. 3) show a steeper gradient than in deciduous teeth, exactly as in modern humans²² but with slightly higher rates at the EDJ (increasing from about 3 μm per day rather than about 2.5 to about 5 μm per day). The total crown formation times in both deciduous and permanent molars were nearly identical with those reported for large samples of modern humans^{15,16}. Protoconid initiation in the M1 occurred about 15 days before birth (neonatal line) and metaconid initiation about 18 days

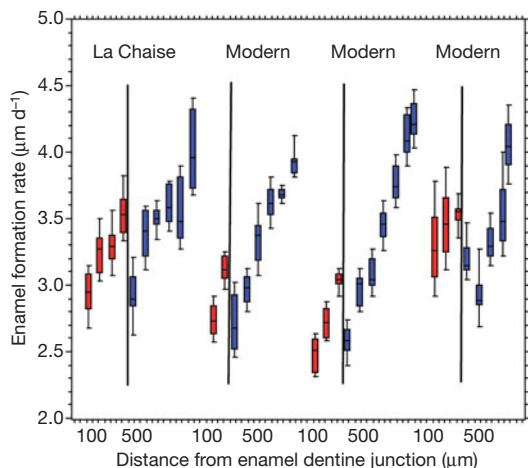


Figure 2 | Graphs of deciduous second molar enamel formation rates for the Neanderthal and three typical modern human deciduous molars. Each box plot is for a minimum of ten groups of five cross-striations averaged at 100- μm zones of enamel as measured from the EDJ (the horizontal lines in each box show 25th, 50th and 75th centiles. The whiskers indicate 10th and 90th centiles). The vertical black line indicates the position of the neonatal line with respect to the EDJ and the measurements of daily secretion rates (red, prenatal; blue, postnatal). The positions of the neonatal line and the recovery phase after birth are similar. In the Neanderthal, the neonatal line occurred about 60 days after initiation in the metaconid and about 100 days in the protoconid. The total crown formation time was about 315 days.

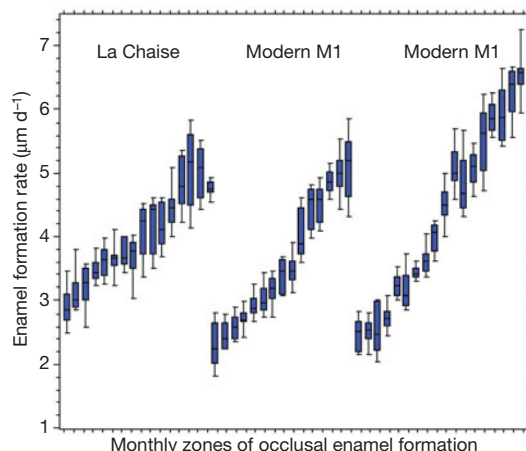


Figure 3 | Graphs of daily occlusal enamel formation rates in the permanent Neanderthal M1 and in modern human molars. Daily rates are averaged as a box plot (with elements defined as in Fig. 2) for successive months of occlusal enamel formation in each molar and are essentially similar. There is a steep gradient of increase in enamel formation rates in all cases. Mean values of measurements in the first monthly zone ($n = 57$ from four modern human molars) were $2.4 \pm 0.3 \mu\text{m}$ per day (mean $\pm$ s.d.). In Neanderthals, rates during the first month were slightly higher and closer to 3 μm per day. Measurements in the M1 from La Chaise ($n = 13$) were $2.9 \pm 0.4 \mu\text{m}$ per day and in Tabun C1 (ref. 18) ($n = 10$) they were $3.2 \pm 0.4 \mu\text{m}$ per day. Maximum rates at the surface enamel increase to about 5 or 6 μm per day in both Neanderthals and modern humans.

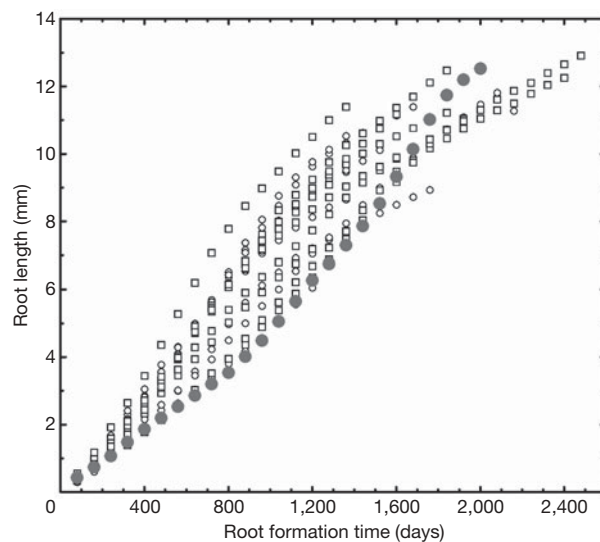


Figure 4 | Growth curves for roots of M1 for 20 modern humans (by sex) and the Neanderthal M1. No differences exist in the distribution of faster or slower growth curves between sexes in modern humans (open circles, females; open squares, males). The Neanderthal root (filled circles) has a slower than average trajectory for the first two-thirds of root growth but then rises quickly towards the end of root growth. The combined time of protoconid formation in La Chaise M1 (1,041 days) and root formation (2,119 days) totals 3,160 days or about 105 months (about 8.7 years). The average extension rate was 6.3 μm per day over a root length of 13.3 mm. The root formation stage in La Chaise M1 was just short of complete closure at the time of death. Estimates of the mean age of attainment of M1 apex formation in modern humans with the 'root canal terminally divergent' (that is, just short of completion) were 100 months, or 8.3 years, in boys (s.d. 7.7 months) and 94.1 months, or 7.8 years, in girls (s.d. 8.1 months) with a mean extension rate for the whole M1 root of 6.8 μm per day (ref. 23). In ref. 23, M1 root formation was, on average, complete at 106.6 months (8.9 years) in boys and 102.5 months (8.54 years) in girls.

after birth. The protoconid took about 1,041 days to form to the cervix and the metaconid about 865 days to the cervix, both only slightly below 1 s.d. of the mean for modern humans of African origin ($1,117 \pm 55$ and 936 ± 55 days). Protoconid lateral enamel formation took about 645 days and expressed about 90 surface perikymata with a periodicity of 7 days.

The timing of root completion in the Neanderthal M1 from La Chaise was reconstructed from incremental markings in the dentine⁷. M1 root completion occurs at about 9 years in modern humans²³ but has not previously been known for Neanderthals. We report here that closure of the root apex was occurring as in modern humans at about 8.7 years of age in this Neanderthal. However, Neanderthal molars are typically taurodont with a long root trunk and with late bifurcation or trifurcation of the roots. Rates of M1 root elongation (root extension²⁴) in modern humans and in this Neanderthal differ (Fig. 4), with slow early root extension rates and a very late peak or spurt shortly after root bifurcation in the Neanderthal (Fig. 5).

Counts and packing patterns of perikymata on anterior tooth surfaces show that the large incisors and canines of Neanderthals formed more quickly than in some modern humans^{4,5}. Periodicities of perikymata determined in this study and from Tabun C1 (ref. 18) (7 and 8 days, respectively) bolster these conclusions. However, gingival emergence of the first permanent molar is a more reliable measure of relative dental development than are anterior tooth crown formation times and is also tightly correlated with brain size in anthropoid primates^{1,2}. Our data now allow us to predict M1 emergence time in Neanderthals with more certainty. If about 8 mm of root were formed at an average of about 5.7 μm per day at gingival emergence, then this would have occurred at about 6.7 years of age, well within the human range (6.2 ± 0.8 years (mean $\pm$ s.d.))²³. This, together with the modern human-like position of the neonatal line, suggests both similar timing of tooth initiation relative to birth in Neanderthals and modern humans, and a predictable extended period between birth and M1 emergence, by which time about 90% of brain volume would have been attained^{1,25}.

One hundred and fifty years after the discovery of the Neanderthal type specimen, the prospects for further work on the complex relationships between jaw size and jaw growth²⁶, early permanent molar initiation^{27,28} and those factors that underlie rates of tooth eruption

that are faster than usual²⁹ are good. Moreover, improved imaging techniques³⁰ combined with further histological analyses of deciduous and permanent tooth enamel are likely to reveal evidence for any changing patterns of perinatal stress as Neanderthals approached extinction and so provide critical evidence about the shift in balance between adult mortality rates, fertility rates and infant survival among more recent Neanderthals than those at La Chaise.

METHODS

Micro-computed tomography analysis. Virtual three-dimensional reconstructions and measurements (Supplementary Information) are based on a high-resolution SR- μCT record performed at the beamline ID17 set at the European Synchrotron Radiation Facility, Grenoble (<http://www.esrf.fr/UsersAndScience/Experiments/Imaging/ID17/>; experiments SC-1587 and SC-1749). The system is characterized by a continuous energy spectrum, a high photon flux, an intense monochromatic X-ray beam, nearly parallel projections and a small angular source size. Samples are located about 140 m from the source point. Monoenergetic X-ray beams enable absolute linear attenuation coefficients to be measured and avoid the risk of beam-hardening artefacts in the reconstruction of images from dense specimens such as highly mineralized fossils³⁰. Scans of the two Neanderthal teeth from La Chaise were performed at energies of 51 keV (deciduous) and 70 keV (permanent). Projections (of 22.6 and 25.6 ms for the deciduous tooth and the permanent tooth, respectively) were taken every 0.12° and were collected by a $2,048 \times 2,048$ fibre-optical taper charge-coupled device Frelon camera. Final sections were reconstructed from sinograms and saved in a 32-bit floating-point raw format at a resolution (voxel size) of $45.5 \times 45.5 \times 45.7 \mu\text{m}^3$. The final eight-bit volumes ($239 \times 262 \times 351$ pixels for the deciduous molar and $493 \times 357 \times 237$ pixels for the permanent molar) were analysed by means of AMIRA v. 4.0 (Mercury Computer Systems, Inc.). Segmentation of the volume was done semi-automatically with manual corrections. The original SR- μCT record is available at the NESPOS website (<https://nespos.pitcom.net/display/openspace/Home>).

Histological analysis. Longitudinal ground sections of molars were prepared in the buccolingual plane. The plane of section passed through the mesial cusps, the last-formed enamel at the mesiobuccal cervix and through the long axis of the root. Often, the apical portion of the root turned distally with respect to this plane of section and was therefore cut obliquely thereafter. Only sections with at least 10 mm of root true to the plane of section were included in this study. An initial 200–300- μm slice was made with a circular diamond saw. Each slice was lapped and polished to a roughly 80- μm -thick section then cleaned and mounted for routine polarized light microscopy. Additional details about calculating root extension rates are provided in Supplementary Information.

Received 20 July; accepted 3 October 2006.

Published online 22 November 2006.

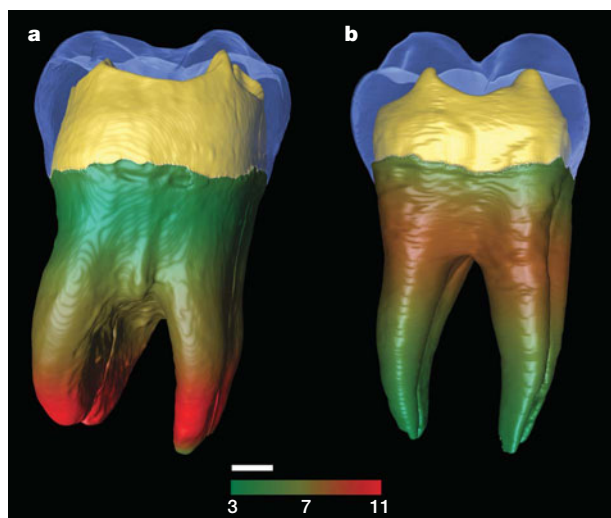


Figure 5 | SR- μCT -based three-dimensional visualization of the differences in root extension rate. a, The Neanderthal M1 from La Chaise. b, An average modern human lower M1. Both are in lingual view. The pseudo-colour scale, ranging from dark green (3 μm per day) to red (11 μm per day), is used to render the topographic variation of the root extension rate from the cervix (yellow–green boundary) towards the radicular apex of the tooth. Extension rate data are from Supplementary Fig. 5. Scale bar, 2 mm.

- Smith, B. H. Dental development and the evolution of life history in Hominidae. *Am. J. Phys. Anthropol.* **86**, 157–174 (1991).
- Smith, B. H. & Tompkins, R. L. Towards a life history of the Hominidae. *Annu. Rev. Anthropol.* **24**, 257–279 (1995).
- Avishai, G. et al. New approach to quantifying developmental variation in the dentition using serial microtomographic imaging. *Microsc. Res. Tech.* **65**, 263–269 (2004).
- Ramirez-Rozzi, F. V. & Bermudez de Castro, J. M. Surprisingly rapid growth in Neanderthals. *Nature* **428**, 936–939 (2004).
- Guatelli-Steinberg, D., Reid, D. J. & Bishop, T. A. Anterior tooth growth periods in Neanderthals were comparable to those in modern humans. *Proc. Natl Acad. Sci. USA* **102**, 14197–14202 (2005).
- Debénath, A. *Néanderthaliens et Cro-Magnons. Les Temps Glaciaires dans le Bassin de la Charente* (Le Croît Vif, Paris, 2006).
- Dean, M. C. Comparative observations on the spacing of short-period (von Ebner's) lines in dentine. *Archs Oral Biol.* **43**, 1009–1021 (1998).
- Dean, M. C. Tooth microstructure tracks the pace of human life-history evolution. *Proc. R. Soc. B* **273**, 2799–2808 (2006).
- Jernvall, J. & Thesleff, I. Reiterative signalling and patterning in mammalian tooth morphogenesis. *Mech. Dev.* **92**, 19–29 (2000).
- Thesleff, I., Keranen, S. & Jernvall, J. Enamel knots as signalling centers linking tooth morphogenesis and odontoblast differentiation. *Adv. Dent. Res.* **15**, 14–18 (2001).
- Bailey, S. E. A closer look at Neanderthal postcanine dental morphology: The mandibular dentition. *Anat. Rec. (New Anat.)* **269**, 148–156 (2002).
- Kono, R. T. Molar enamel thickness and distribution patterns in extant great apes and humans: New insights based on a 3-dimensional whole crown perspective. *Anthropol. Sci.* **112**, 121–146 (2004).
- Olejniczak, A., Martin, L. B. & Ulhaas, L. Quantification of dentine shape in anthropoid primates. *Ann. Anat.* **186**, 479–485 (2004).

14. Suwa, G. & Kono, R. T. A micro-CT based study of linear enamel thickness in the mesial cusp section of human molars: Reevaluation of methodology and assessment of within-tooth, serial, and individual variation. *Anthropol. Sci.* **113**, 273–289 (2005).
15. Liversidge, H. M. & Molleson, T. Variation in crown and root formation and eruption of human deciduous teeth. *Am. J. Phys. Anthropol.* **123**, 172–180 (2004).
16. Reid, D. J. & Dean, M. C. Variation in modern human enamel formation times. *J. Hum. Evol.* **50**, 329–346 (2006).
17. Dean, M. C., Stringer, C. B. & Bromage, T. G. Age at death of the Neanderthal child from Devils Tower, Gibraltar and the implications for studies of general growth and development in Neanderthals. *Am. J. Phys. Anthropol.* **70**, 301–309 (1986).
18. Stringer, C. B., Dean, M. C. & Martin, R. D. in *Primate Life History and Evolution: Monographs in Primatology* Vol. 14 (ed. De Rousseau, C. J.) 115–152 (Wiley-Liss, New York, 1990).
19. Trinkaus, E. & Tompkins, R. L. in *Primate Life History and Evolution: Monographs in Primatology* Vol. 14 (ed. De Rousseau, C. J.) 153–180 (Wiley-Liss, New York, 1990).
20. Tillier, A.-M., Mann, A. E., Monge, J. & Lampl, M. L'ontogenèse, la croissance de l'émail dentaire et l'origine de l'homme moderne: l'exemple des Néandertaliens. *Anthropol. Préhist.* **106**, 97–104 (1995).
21. Rossi, P. F., Bondioli, L., Geusa, G. & Macchiarelli, R. in *Digital Archives of Human Paleobiology* (eds Bondioli, L. & Macchiarelli, R.) (E-LISA, Milano, 1999) [CD-ROM].
22. Dean, M. C. *et al.* Growth processes in teeth distinguish modern humans from *Homo erectus* and earlier hominins. *Nature* **414**, 628–631 (2001).
23. Gleiser, I. & Hunt, E. E. The permanent mandibular molar: Its calcification, eruption and decay. *Am. J. Phys. Anthropol.* **13**, 253–283 (1955).
24. Shellis, R. P. Variations in growth of the enamel crown in human teeth and a possible relationship between growth and enamel structure. *Archs Oral Biol.* **29**, 697–705 (1984).
25. Ashton, E. H. & Spence, T. F. Age changes in the cranial capacity and foramen magnum of hominoids. *Proc. Zool. Soc. Lond.* **130**, 169–180 (1958).
26. Kondo, O. *et al.* in *Current Trends in Dental Morphology Research* (ed. E. Żądzińska) 243–255 (Univ. Lodz Press, 2005).
27. Boughner, J. C. & Dean, M. C. Does space in the jaw influence the timing of molar crown initiation? A model using baboons and great apes. *J. Hum. Evol.* **46**, 253–275 (2004).
28. Tompkins, R. L. Relative dental development of Upper Pleistocene hominids compared to human population variability. *Am. J. Phys. Anthropol.* **99**, 103–118 (1996).
29. Wolpoff, M. H. The Krapina dental remains. *Am. J. Phys. Anthropol.* **50**, 67–114 (1979).
30. Mazurier, A., Volpato, V. & Macchiarelli, R. Improved noninvasive microstructural analysis of fossil tissues by means of SR-microtomography. *Appl. Phys. A* **83**, 229–233 (2006).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Bravin, C. Nemoz and P. Tafforeau for collaboration at the ESRF beamline ID17; the Centre de Microtomographie (CdMT) at the University of Poitiers; the Department of Physics at the University of Bologna; and P. Bayle, A. Bergeret, P. Sardini, V. Volpato and P. Walton for technical assistance. The research was supported by the French CNRS, the EU TNT Project (to R.M.), the Région Poitou-Charentes (to A.M.), and The Leverhulme Trust and The Royal Society (to C.D.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.D. (ucgacrd@ucl.ac.uk).

LETTERS

Climate-driven trends in contemporary ocean productivity

Michael J. Behrenfeld¹, Robert T. O'Malley¹, David A. Siegel³, Charles R. McClain⁴, Jorge L. Sarmiento⁵, Gene C. Feldman⁴, Allen J. Milligan¹, Paul G. Falkowski⁶, Ricardo M. Letelier² & Emmanuel S. Boss⁷

Contributing roughly half of the biosphere's net primary production (NPP)^{1,2}, photosynthesis by oceanic phytoplankton is a vital link in the cycling of carbon between living and inorganic stocks. Each day, more than a hundred million tons of carbon in the form of CO₂ are fixed into organic material by these ubiquitous, microscopic plants of the upper ocean, and each day a similar amount of organic carbon is transferred into marine ecosystems by sinking and grazing. The distribution of phytoplankton biomass and NPP is defined by the availability of light and nutrients (nitrogen, phosphate, iron). These growth-limiting factors are in turn regulated by physical processes of ocean circulation, mixed-layer dynamics, upwelling, atmospheric dust deposition, and the solar cycle. Satellite measurements of ocean colour provide a means of quantifying ocean productivity on a global scale and linking its variability to environmental factors. Here we describe global ocean NPP changes detected from space over the past decade. The period is dominated by an initial increase in NPP of 1,930 teragrams of carbon a year (Tg C yr⁻¹), followed by a prolonged decrease averaging 190 Tg C yr⁻¹. These trends are driven by changes occurring in the expansive stratified low-latitude oceans and are tightly coupled to coincident climate variability. This link between the physical environment and ocean biology functions through changes in upper-ocean temperature and stratification, which influence the availability of nutrients for phytoplankton growth. The observed reductions in ocean productivity during the recent post-1999 warming period provide insight on how future climate change can alter marine food webs.

Interannual changes in phytoplankton abundance are small relative to total standing stocks and require accurate, long-term measurements to detect. The first satellite sensor to have this capability is the Sea-viewing Wide Field-of-View Sensor (SeaWiFS) and so far it is the only ocean colour sensor to have lunar-viewing calibration capabilities³. These measurements allow SeaWiFS to achieve water-leaving radiance biases across the SeaWiFS visible bands (412–555 nm) of less than 4% when compared to field open-ocean observations⁴. The central biological variable derived from these measurements is upper-ocean chlorophyll concentration, which can be used to estimate phytoplankton standing stocks and productivity (Fig. 1a) throughout the photic zone⁵.

Global, depth-integrated chlorophyll biomass (ΣChl) for the 1997 to 2006 SeaWiFS record varied from 4.2 to 5.0 Tg (Tg = 10¹² g). Monthly anomalies in global ΣChl for this period reveal two highly significant ($P < 0.0001$) trends (green line in Fig. 1b). The initial 0.3 Tg rise in ΣChl corresponds to the 1997 to 1999 El Niño to La Niña transition². Since 1999, global chlorophyll concentrations

dropped on average 0.01 Tg per year, with several secondary features that ended in a 2005 to 2006 rise (Fig. 1b). The two primary trends in global ΣChl are driven by changes occurring in the permanently stratified regions of the ocean (grey circles in Fig. 1b). These regions have annual average sea surface temperatures (SSTs) over 15 °C (black lines in Fig. 1a) and represent 74% of the global ice-free ocean. Cooler, seasonal seas also experienced local variability during the SeaWiFS period, but when integrated spatially they exhibit no significant overall trend ($P > 0.5$) (Supplementary Fig. 1).

The entire phytoplankton biomass of the global oceans is consumed every two to six days⁶. This rapid turnover implies that variability in ΣChl (Fig. 1b) corresponds to far more substantial changes in NPP. One of the most commonly used models for estimating NPP from satellite chlorophyll data is the Vertically Generalized Production Model (VGPM)⁶. The VGPM gives global NPP estimates for the SeaWiFS record that vary strongly with season and range from 3.8 to 4.6 × 10¹⁵ g of C per month. Monthly anomalies in this important carbon flux again reveal clear ($P < 0.0001$) temporal trends beginning with an El Niño/La Niña-driven 262 Tg increase in NPP between 1997 and 1999, followed by a 197 Tg decrease to 2004, and finally a modest increase during 2005 (green line in Fig. 1c). These prominent NPP trends are again traceable to changes in the permanently stratified oceans (grey circles in Fig. 1c), are highly correlated ($r^2 = 0.93$) with changes in ΣChl (Fig. 1b), and remain when NPP is reassessed using alternative productivity models (Supplementary Fig. 2).

Perhaps the most challenging aspect of understanding variability in biological processes is associating detected changes with the environmental forcings responsible. Important climate variables linked to ocean circulation and productivity are sea-level pressure, surface winds, SST, surface air temperature, and cloudiness. These physical factors have been combined into a Multivariate ENSO Index (MEI) for evaluating the strength of El Niño/Southern Oscillation cycles^{7,8}. Strong El Niño/Southern Oscillation events have major impacts on phytoplankton, fisheries, marine birds and mammals^{2,9–12} and are striking examples of climatic influences on ocean biology. Importantly, the MEI does not distinguish between natural and anthropogenic changes in climate forcing, but instead provides an integrated index of climate conditions for comparison with observed changes in ocean productivity.

We find a clear, strong correspondence between MEI variability and SeaWiFS-based anomalies in NPP ($r^2 = 0.77$, $P < 0.005$) (Fig. 2a) and ΣChl (Supplementary Fig. 3). An increase in the MEI (that is, warmer conditions) results in a decrease in NPP and ΣChl , and vice versa. This relationship emphasizes the pre-eminent role of climate

¹Department of Botany and Plant Pathology, 2082 Cordley Hall, and ²College of Oceanographic and Atmospheric Sciences, Oregon State University, Corvallis, Oregon 97331, USA.

³Institute for Computational Earth System Science and Department of Geography, University of California, Santa Barbara, Santa Barbara, California 93106-3060, USA. ⁴NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA. ⁵Atmospheric and Oceanic Sciences Program, Princeton University, PO Box CN710, Princeton, New Jersey 08544, USA. ⁶Environmental Biophysics and Molecular Ecology Program, Institute of Marine and Coastal Sciences and Department of Geological Sciences, Rutgers University 71 Dudley Rd, New Brunswick, New Jersey 08901, USA. ⁷School of Marine Sciences, 209 Libby Hall, University of Maine, Orono, Maine 04469-5741, USA.

variability on ocean productivity trends. The observed physical–biological coupling between NPP and the MEI (Fig. 2a) functions through an effect of climate on water column stratification. Climatic changes that allow surface warming cause an increase in the density contrast between the surface layer and underlying nutrient-rich waters. For the global stratified oceans, the density difference between the surface and a depth of 200 m provides a useful measure of stratification. Monthly anomalies in this stratification index exhibit similar trends as the MEI for the SeaWiFS period (Fig. 2b). Importantly, enhanced stratification suppresses nutrient exchange through vertical mixing. Conversely, surface cooling favours elevated vertical exchange. Phytoplankton in the ocean's upper layer (that is, the populations observed from space) rely on vertical nutrient transport to sustain productivity, so intensified stratification during a rising MEI period (Fig. 2b) is accompanied closely by decreasing NPP (Fig. 2b) ($r^2 = 0.73$, $P < 0.005$).

The MEI captures global-scale trends in climate forcings on surface ocean growth conditions for phytoplankton (Fig. 2). However, while an increasing MEI corresponds to an overall warming period, at the

regional scale both warming and cooling events are found. These regional changes are registered in satellite SST data and are well illustrated by the 1999 to 2004 period of increasing MEI. Comparison of SST changes (Fig. 3a) and modelled NPP changes (Fig. 3b) for this period reveals the anticipated inverse relationship of increasing temperatures coupled to decreasing production (Fig. 3c). Quantitatively, over 74% of the global oceans with SST changes exceeding $\pm 0.15^\circ\text{C}$ between 1999 and 2004 exhibited an inverse relationship between temperature and NPP changes (Fig. 3c). A consistent relationship is likewise observed between SST and ΣChl (Supplementary Fig. 4).

Our results clearly link ΣChl and NPP changes to SST, stratification and climate variations and provide an opportunity for comparison with simulated ocean responses to a warming climate. Atmosphere–ocean general circulation models coupled to ocean ecosystem models are essential tools for understanding global carbon cycle–climate feedbacks. When forced by rising atmospheric CO_2 concentrations, these prognostic models consistently yield increased SSTs and net decreases in stratified ocean NPP^{13–16}, consistent with observed changes in NPP during the 1999 to 2004 period of rising MEI (Fig. 2a). Also consistent with our results, model-simulated NPP changes are largely due to intensified water-column stratification and are not uniformly distributed across the stratified oceans^{13–16}. A warming climate in atmosphere–ocean general circulation models is also generally associated with increased NPP at higher latitudes, owing to improved mixed-layer light conditions and extended growing seasons^{14–16}. For our relatively brief observational record, this anticipated net increase in high-latitude NPP was not observed, although NPP changes were found regionally (Fig. 3b).

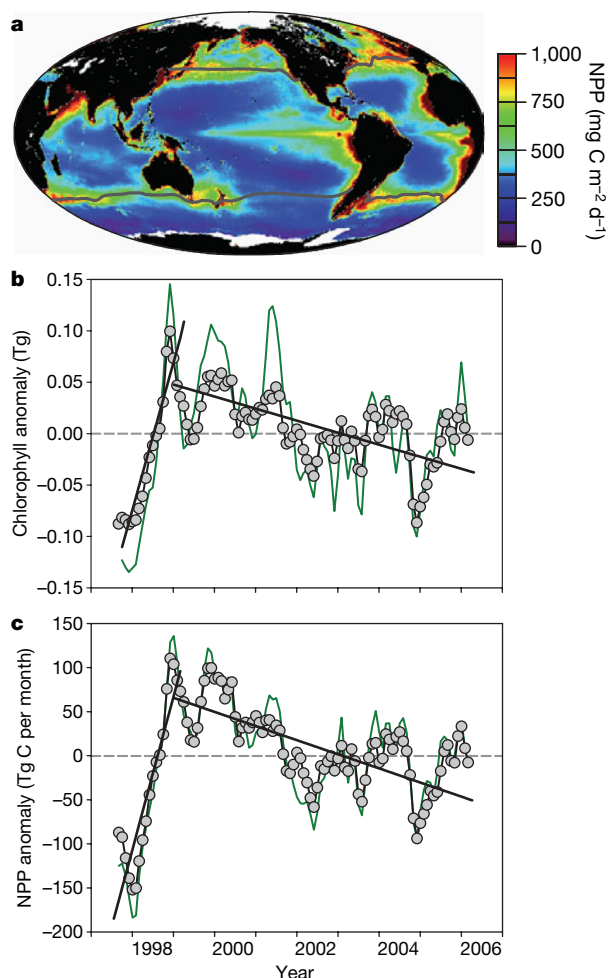


Figure 1 | Distribution and trends in global ocean phytoplankton productivity (NPP) and chlorophyll standing stocks. **a**, Annual average NPP showing high values where surface nutrients are elevated. Low-latitude, permanently stratified waters with annual average surface temperatures over 15°C are delineated by black contour lines. **b**, Anomalies in globally integrated water-column chlorophyll concentrations (green line) are dominated by changes occurring in permanently stratified ocean regions (grey circles and black line). **c**, Anomalies in global NPP (green line) are likewise driven by changes in the permanently stratified oceans (grey circles and black line). Trend lines in **b** and **c** are least-squares fits to pre-1999 and post-1999 data.

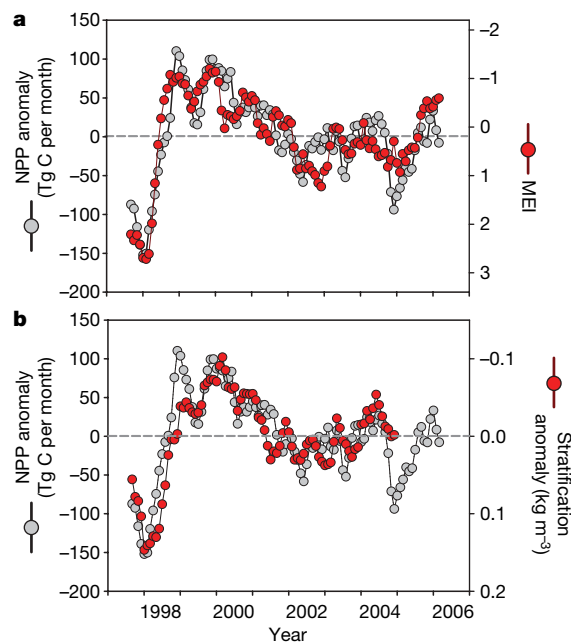


Figure 2 | Ocean productivity is closely coupled to climate variability. **a**, NPP anomalies in the permanently stratified oceans (grey symbols, left axis) are highly correlated ($r^2 = 0.77$) with the MEI of climate variability (red symbols, right axis). NPP data are from Fig. 1c. **b**, Changes in ocean stratification (red symbols, right axis) link climate variability to ocean biology, and are well correlated ($r^2 = 0.73$) with NPP anomalies (grey symbols, left axis) in ocean regions with annual average surface temperatures over 15°C . Stratification strength was assessed as the density differences between the surface and a depth of 200 m using SODA data (see Methods). Publicly accessible SODA data are available only to the end of 2004. Note that the MEI and stratification axes (right) increase from top to bottom.

Establishing relationships between ocean biology and climate relies heavily on the accuracy to which changes in ocean ecosystems can be detected³. The unique operational characteristics of SeaWiFS have enabled detection of the global ocean trends reported here. Unfortunately, the SeaWiFS sensor is well beyond its design lifetime and no currently scheduled ocean-colour missions will have equal capabilities. SeaWiFS also represents an immature state of ocean-colour remote sensing. Modelling studies suggest that shifts in ecosystem structure from climate variations may be as or more important than the alterations in bulk integrated properties reported here¹³. Susceptible ecosystem characteristics include shifts in taxonomic composition¹³, physiological status¹⁷, and light absorption by coloured dissolved organic material¹⁸. Resolving these properties and their changes over time will require improved accuracies and significant expansions in the spectral range and resolution of future remote-sensing measurements.

Climate effects on ocean biology are documented here for nearly a decade of satellite ocean colour measurements. The compelling correspondence found between NPP changes and the MEI (Fig. 2a) suggests that this integrated index of climate variability, which extends back to 1950 (refs 6, 7), may provide a first-order proxy for variations in photosynthetic carbon fixation in the oceans for the period predating satellite ocean-colour measurements. It is also clear from the current analysis that surface warming in the permanently stratified ocean regions is accompanied by reductions in productivity. The index used here (MEI) to relate climate variability to NPP trends does not distinguish natural from anthropogenic contributions, but observational and modelling efforts indicate that recent changes in SST are strongly influenced by anthropogenic forcing^{19–22}.

These observations imply that the potential transition to permanent El Niño conditions in a warmer climate state²³ would lead to lower and redistributed ocean carbon fixation relative to typical contemporary conditions. Such changes will inevitably alter the magnitude and distribution of global ocean net air–sea CO₂ exchange^{15,24,25}, fishery yields^{9,12,26}, and dominant basin-scale biological regimes¹².

METHODS

Ocean NPP was calculated with the VGPM using monthly 1,080 × 2,160 pixel resolution (that is, 18 km spacing at the Equator) OC4-v4 chlorophyll algorithm products from SeaWiFS reprocessed version 5.1 data (<http://oceancolor.gsfc.nasa.gov>). Comparison of this data with ~1,400 *in situ* match-up surface chlorophyll data yields a median difference of 33%, which is comparable to measurement uncertainties in the field. ΣChl was calculated from the satellite chlorophyll data using relationships described in ref. 5, which are based on 3,806 *in situ* chlorophyll profiles and account for subsurface chlorophyll features. The VGPM describes light-dependent changes in water-column NPP using a relationship based on 1,698 field productivity measurements⁶ and SeaWiFS cloud-corrected photosynthetically active radiation data (<http://oceancolor.gsfc.nasa.gov>). The standard VGPM describes photosynthetic efficiencies using a polynomial function that increases with temperature below 20 °C and decreases above 20 °C.

Ocean NPP was additionally calculated using two alternative models, with results demonstrating the robust nature of the primary temporal trends described here (Supplementary Fig. 2). MEI data were from <http://www.cdc.noaa.gov/people/klaus.wolter/MEI>. Stratification anomalies were calculated from Simple Ocean Data Assimilation (SODA) (<http://www.atmos.umd.edu/~ocean>) monthly global density data, which assimilate available field observations. AVHRR SST data were from <http://podaac-www.jpl.nasa.gov/sst>. Deseasonalized chlorophyll, NPP, and stratification anomalies were calculated by subtracting from each monthly value the corresponding monthly average for the entire time series. Reported statistics for relationships between NPP anomalies and MEI and stratification anomalies account for auto-correlation effects inherent in comparisons between time-series data sets.

Received 18 August; accepted 6 October 2006.

- Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. G. Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**, 237–240 (1998).
- Behrenfeld, M. J. *et al.* Biospheric primary production during an ENSO transition. *Science* **291**, 2594–2597 (2001).
- McClain, C. R., Feldman, G. C. & Hooker, S. B. An overview of the SeaWiFS project and strategies for producing a climate research quality global ocean bio-optical time series. *Deep-Sea Res.* **51**, 5–42 (2004).
- Bailey, S. W. & Werdell, P. J. A multi-sensor approach for the on-orbit validation of ocean color satellite data products. *Remote Sens. Environ.* **102**, 12–23 (2006).
- Morel, A. & Berthon, J.-F. Surface pigments, algal biomass profiles, and potential production of the euphotic layer: relationships reinvestigated in view of remote-sensing applications. *Limnol. Oceanogr.* **34**, 1545–1562 (1989).
- Behrenfeld, M. J. & Falkowski, P. G. Photosynthetic rates derived from satellite-based chlorophyll concentration. *Limnol. Oceanogr.* **42**, 1–20 (1997).
- Wolter, K. The Southern Oscillation in surface circulation and climate over the tropical Atlantic, Eastern Pacific, and Indian Oceans as captured by cluster analysis. *J. Clim. Appl. Meteorol.* **26**, 540–558 (1987).
- Wolter, K. & Timlin, M. S. Monitoring ENSO in COADS with a seasonally adjusted principal component index. In *Proc. 17th Climate Diagnostics Workshop (Norman, Oklahoma)* 52–57 (NOAA/N MC/CAC, NSSL, Oklahoma Climate Survey, CIMMS and the School of Meteorology, Univ. Oklahoma, 1993).
- Barber, R. T. & Chavez, F. P. Biological consequences of El Niño. *Science* **222**, 1203–1210 (1983).
- Chavez, F. P. *et al.* Biological and chemical response of the equatorial Pacific ocean to the 1997–98 El Niño. *Science* **286**, 2126–2131 (1999).
- Turk, D., McPhaden, M. J., Lewis, M. R. & Busalacchi, A. J. Remotely-sensed biological production in the Tropical Pacific during 1992–1999. *Science* **293**, 471–474 (2001).
- Chavez, F. P., Ryan, J., Lluch-Cota, S. E. & Niquen, M. C. From anchovies to sardines and back: Multidecadal change in the Pacific ocean. *Science* **299**, 217–221 (2003).
- Boyd, P. W. & Doney, S. C. Modelling regional responses by marine pelagic ecosystems to global climate change. *Geophys. Res. Lett.* **29**, 1806, doi:10.1029/2001GL014130 (2002).
- Bopp, L. *et al.* Potential impact of climate change on marine export production. *Glob. Biogeochem. Cycles* **15**, 81–99 (2001).
- Le Quéré, C., Aumont, O., Monfray, P. & Orr, J. Propagation of climatic events on ocean stratification, marine biology, and CO₂: Case studies over the

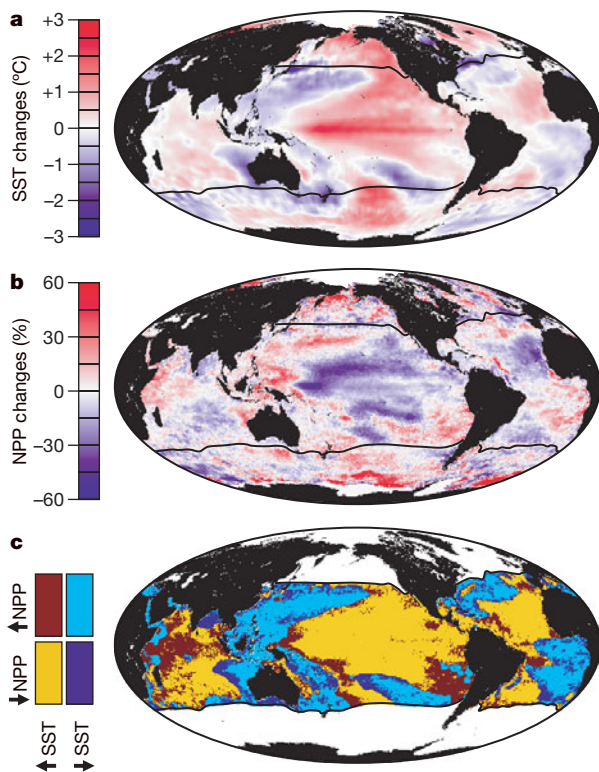


Figure 3 | Climate controls on ocean productivity cause NPP to vary inversely with changes in SST. Global changes in annual average SST (a) and NPP (b) for the 1999 to 2004 warming period (Fig. 2). c, For 74% of the permanently stratified oceans (that is, regions between black contour lines), NPP and SST changes were inversely related. Yellow, increase in SST, decrease in NPP. Light blue, decrease in SST, increase in NPP. Dark blue, decreases in SST and NPP. Dark red, increases in SST and NPP. A similar inverse relationship is observed between SST and chlorophyll changes. (See Supplementary Fig. 4 for additional information.)

- 1979–1999 period. *J. Geophys. Res.* **108**, 3375, doi:10.1029/2001JC000920 (2003).
16. Sarmiento, J. L. *et al.* Response of ocean ecosystems to climate warming. *Glob. Biogeochem. Cycles* **18**, GB3003, doi:10.1029/2003GB002134 (2004).
 17. Behrenfeld, M. J., Boss, E., Siegel, D. A. & Shea, D. M. Carbon-based ocean productivity and phytoplankton physiology from space. *Glob. Biogeochem. Cycles* **19**, GB1006, doi:10.1029/2004GB002299 (2005).
 18. Siegel, D. A., Maritorena, S., Nelson, N. B. & Behrenfeld, M. J. Independence and interdependences of global ocean optical properties viewed using satellite color imagery. *J. Geophys. Res.* **110**, C07011, doi:10.1029/2004JC002527 (2005).
 19. Levitus, S., Antonov, J. I., Boyer, T. P. & Stephens, C. Warming of the world ocean. *Science* **287**, 2225–2229 (2000).
 20. Barnett, T. P., Pierce, D. W. & Schnur, R. Detection of anthropogenic climate change in the world's oceans. *Science* **292**, 270–274 (2001).
 21. Levitus, S. *et al.* Anthropogenic warming of Earth's climate system. *Science* **292**, 267–270 (2001).
 22. Barnett, T. P. *et al.* Penetration of human-induced warming into the world's oceans. *Science* **309**, 284–287 (2005).
 23. Wara, M. W., Ravelo, A. C. & DeLaney, M. L. Permanent El Niño-like conditions during the Pliocene warm period. *Science* **309**, 758–761 (2005).
 24. Sarmiento, J. L., Hughes, T. M. C., Stouffer, R. J. & Manabe, S. Simulated response of the ocean carbon cycle to anthropogenic climate warming. *Nature* **393**, 245–249 (1998).
 25. Laws, E. A., Falkowski, P. G., Smith, W. O. Jr & McCarthy, J. J. Temperature effects on export production in the open ocean. *Glob. Biogeochem. Cycles* **14**, 1231–1246 (2000).
 26. Ware, D. M. & Thomson, R. E. Bottom-up ecosystem trophic dynamics determine fish production in the Northeast Pacific. *Science* **308**, 1280–1284 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the Ocean Biology Processing Group at the Goddard Space Flight Center, Greenbelt, Maryland for their diligence in providing the highest quality ocean colour products possible and the NASA Ocean Biology and Biogeochemistry Program for support.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.J.B. (mjb@science.oregonstate.edu).

LETTERS

Glioma stem cells promote radioresistance by preferential activation of the DNA damage response

Shideng Bao^{1,2}, Qiulian Wu^{1,2}, Roger E. McLendon^{2,3}, Yueling Hao^{1,2}, Qing Shi^{1,2}, Anita B. Hjelmeland^{1,2}, Mark W. Dewhirst⁴, Darell D. Bigner^{2,3} & Jeremy N. Rich^{1,2,5,6}

Ionizing radiation represents the most effective therapy for glioblastoma (World Health Organization grade IV glioma), one of the most lethal human malignancies¹, but radiotherapy remains only palliative² because of radioresistance. The mechanisms underlying tumour radioresistance have remained elusive. Here we show that cancer stem cells contribute to glioma radioresistance through preferential activation of the DNA damage checkpoint response and an increase in DNA repair capacity. The fraction of tumour cells expressing CD133 (Prominin-1), a marker for both neural stem cells and brain cancer stem cells^{3–6}, is enriched after radiation in gliomas. In both cell culture and the brains of immunocompromised mice, CD133-expressing glioma cells survive ionizing radiation in increased proportions relative to most tumour cells, which lack CD133. CD133-expressing tumour cells isolated from both human glioma xenografts and primary patient glioblastoma specimens preferentially activate the DNA damage checkpoint in response to radiation, and repair radiation-induced DNA damage more effectively than CD133-negative tumour cells. In addition, the radioresistance of CD133-positive glioma stem cells can be reversed with a specific inhibitor of the Chk1 and Chk2 checkpoint kinases. Our results suggest that CD133-positive tumour cells represent the cellular population that confers glioma radioresistance and could be the source of tumour recurrence after radiation. Targeting DNA damage checkpoint response in cancer stem cells may overcome this radioresistance and provide a therapeutic model for malignant brain cancers.

Glioblastomas are the most lethal primary brain tumour with a median survival of less than 12 months because of resistance to radiation and other treatments¹. Glioblastomas present as diffuse tumours with invasion into normal brain, but frequently recur or progress after radiation as focal masses², suggesting that only a fraction of tumour cells is responsible for regrowth. Identification of a crucial cellular subpopulation of brain tumour cells with potent tumorigenic activity^{3–5,7} supports the cancer stem cell hypothesis in solid tumours. As glioma subpopulations expressing Prominin-1 (CD133⁺) are enriched for cancer stem cells and show greater tumorigenic potential than do CD133[−] cells^{3–5}, we have examined the role of glioma cancer stem cells in the development of radioresistance.

We found that ionizing radiation (IR) treatment of short-term cultures from human glioma xenografts enriched the CD133⁺ subpopulation fourfold relative to untreated cultures (Fig. 1a, b). Likewise, glioma xenografts irradiated *in vivo* were enriched 3–5-fold for CD133⁺ cells relative to untreated xenografts (Fig. 1c). Similar results were seen in freshly isolated glioblastoma tumour specimens: the basal fraction of CD133⁺ cells was 2–3% and increased to 6–10% after IR treatment (Fig. 1d). Irradiation did not induce CD133

expression in CD133[−] tumour cells (Supplementary Fig. S1), confirming that increased CD133⁺ fractions after IR were caused by enrichment of original CD133⁺ subpopulations. In addition, two radioresistant human glioma cell subpopulations derived from short-term cultures of glioma xenografts subjected to three serial cycles of IR also contained greater percentages of CD133⁺ cells than parental populations (Supplementary Fig. S2). Thus, tumours surviving IR are enriched in CD133⁺ cancer cells.

To define the biological significance of CD133⁺ enrichment after IR, we implanted a constant number of tumour cells with increasing percentages of CD133⁺ cells into the frontal lobes of immunocompromised mice. Increased CD133⁺ cell fractions dose-dependently decreased tumour latency and enhanced tumour growth and vascularity (Fig. 1e, f, and data not shown). Consistent with these results, viable tumour cells from irradiated xenografts were enriched for CD133⁺ cells and formed secondary tumours with decreased latencies relative to untreated xenografts (Supplementary Fig. S3). Thus, enrichment of CD133⁺ cells is crucial in glioma recurrence after radiotherapy.

To confirm that purified CD133⁺ subpopulations are enriched for cancer stem cells, we characterized CD133⁺ and CD133[−] cells derived from glioma xenografts and primary human glioblastoma samples. Murine host contamination of CD133[−] cellular subpopulations from xenografts was ruled out with the pan-human antibody 3B4, which binds to a glycoprotein only on human cells⁸ (Supplementary Fig. S4). The neoplastic origin of cells derived from the human biopsy specimens was confirmed through fluorescent *in situ* hybridization (FISH) analysis of genetic markers altered in the original patient tumour specimen (Supplementary Fig. S5 and data not shown). CD133⁺ tumour cells showed characteristics consistent with cancer stem cells^{3–7,9–11}: namely, neurosphere formation (Fig. 2a); expression of neural and/or cancer stem cell markers, including CD133, Sox2, Musashi and Nestin (Fig. 2b); and multilineage differentiation with markers for astrocytes (GFAP, S100 β), neurons (Map-2, TUJ1) or oligodendrocytes (O4, GalC) (Supplementary Figs S6 and S7a). CD133⁺ cells derived from xenografts or biopsy specimens formed neurospheres (76–89%; Supplementary Table 1), whereas CD133[−] cells rarely formed neurospheres. CD133⁺ tumour cells were highly tumorigenic in brains of immunocompromised mice with characteristics of glioblastomas (Fig. 2c–e and Supplementary Table 1) in concordance with previous reports^{3–5}. CD133[−] cells did not form detectable tumours even when implanted at 2×10^6 cells per mouse, except for occasional tumours from a single xenograft source (D456MG paediatric xenografts; Fig. 2c and Supplementary Table 1). Of note, we used a shorter incubation period in our mouse studies than some other researchers^{3,7}. Longer incubation periods may permit greater regrowth of poorly

¹Department of Surgery, ²Preston Robert Tisch Brain Tumor Center, ³Department of Pathology, ⁴Department of Radiation Oncology, ⁵Department of Medicine, and ⁶Department of Neurobiology, Duke University Medical Center, Durham, North Carolina 27710, USA.

tumorigenic cells but increases the probability of secondary genetic changes. Thus, CD133⁺ subpopulations derived from patient glioma specimens and xenografts are enriched for characteristics of cancer stem cells, including tumorigenesis *in vivo*.

To identify further the cell subpopulations that contribute to glioma radioresistance, we studied the radiosensitivity of CD133⁺ and CD133⁻ tumour cell subpopulations. Colony formation assays confirmed that CD133⁺ cells isolated from xenografts and a biopsy specimen were more resistant to IR treatment than were corresponding CD133⁻ cells (Fig. 3a and Supplementary Fig. S8a). The differential radioresistance between the two subpopulations was consistent regardless of the presence of growth factors (Supplementary Figs S9 and S10). The preferential survival of CD133⁺ cells after irradiation was due to lower rates of apoptosis, as indicated by decreased activation of caspase-3 in CD133⁺ cells from irradiated cultures or xenografts (Fig. 3b and Supplementary Fig. S8b). In addition, annexin V

staining showed that IR-induced apoptosis in CD133⁺ cells was 4–5-fold lower than that in matched CD133⁻ cells isolated from primary glioblastomas (Supplementary Fig. S11).

To verify the preferential survival of CD133⁺ cells after IR, matched CD133⁺ and CD133⁻ cells from a xenograft or patient biopsy specimen were differentially labelled with fluorescent dyes^{12,13} and mixed in defined ratios. In the absence of IR, the relative percentage of cells derived from CD133⁺ cells increased only modestly over time. By contrast, the percentage of CD133⁺-derived cells after irradiation increased more than fourfold (Fig. 3c, d, and Supplementary Fig. S12), confirming that CD133⁺ tumour cells have greater radioresistance and repopulation potential than do CD133⁻ cells *in vitro*.

We examined the capacity of CD133⁺ glioma cells to form tumours after irradiation in several assays. In an *in vivo* limiting dilution tumour formation assay, irradiated CD133⁺ cells from a

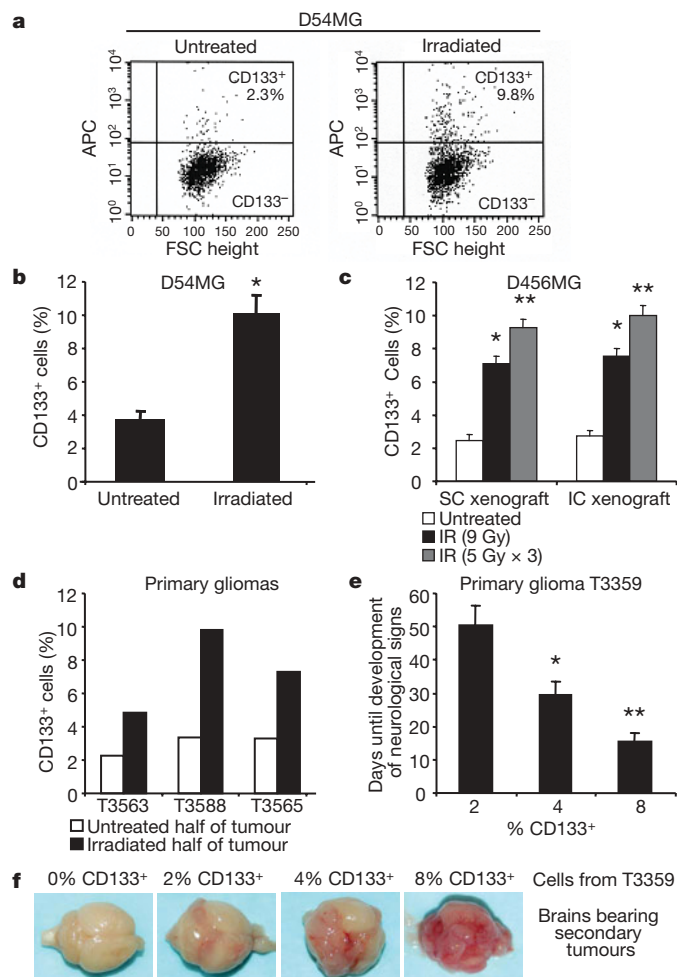


Figure 1 | Enrichment of CD133⁺ tumour subpopulations after irradiation *in vitro* and *in vivo*, and enhancement of intracranial tumour formation by increased CD133⁺ fraction. **a**, Glioma cultures from D54MG xenografts were untreated or irradiated (5 Gy). CD133⁺ fractions were assayed by FACS after 48 h. **b**, Mean $\pm$ s.d. results from **a** ($n = 3$; * $P < 0.001$). **c**, Subcutaneous (SC) and intracranial (IC) D456MG xenografts were irradiated (3×5 Gy or 1×9 Gy) or untreated. CD133⁺ fractions were quantified after 48 h (mean $\pm$ s.d., $n = 3$; * $P < 0.002$; ** $P < 0.001$). **d**, Individual patient tumour specimens were halved and either irradiated with 2 Gy or untreated. The CD133⁺ fraction in each sample was quantified by FACS. **e**, **f**, CD133⁺ and CD133⁻ cells from glioblastoma specimen T3359 were mixed in different ratios and xenotransplanted into mouse brains (100,000 total cells per mouse). **e**, Survival until development of neurological signs (mean $\pm$ s.d., $n = 5$; * $P < 0.002$; ** $P < 0.001$). **f**, Representative images of brains bearing secondary tumours.

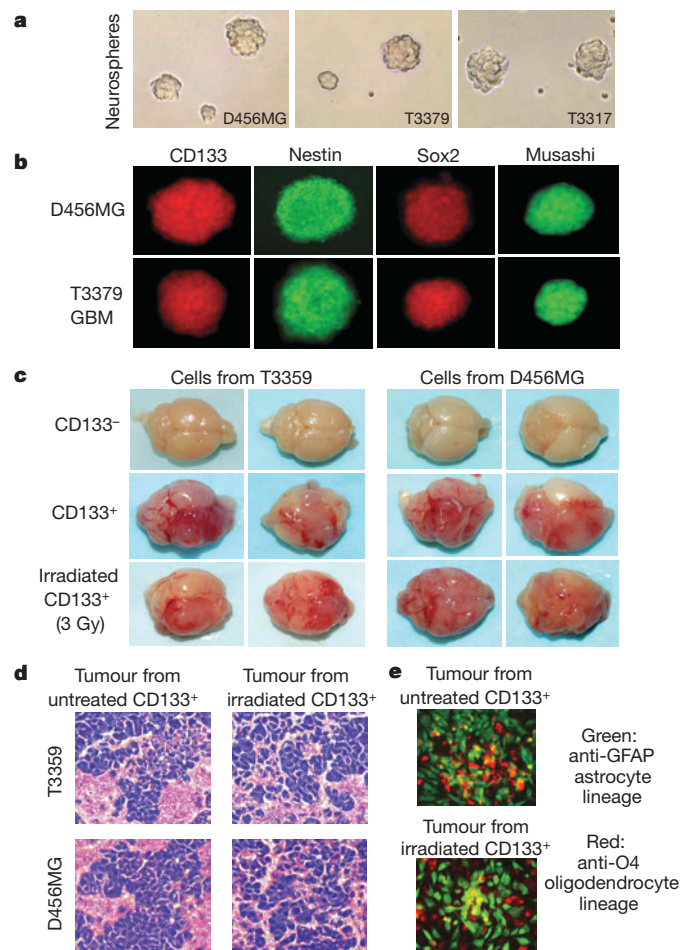


Figure 2 | Characterization of CD133⁺ and CD133⁻ cells from human glioma xenografts and primary glioblastoma specimens. **a**, Representative images of neurospheres from CD133⁺ cells from D456MG xenografts and glioblastoma specimens (T3379, T3317). **b**, Neurospheres from **a** expressed neural stem cell markers (CD133, Nestin, Sox2 and Musashi), as assessed by immunofluorescence. **c**, Untreated or irradiated (3 Gy) CD133⁺ cells (10^4) from glioblastoma specimen T3359 and from D456MG xenografts were transplanted into brains of immunocompromised mice ($n = 5$). Mice were killed on development of neurological signs or after 8 weeks. CD133⁺ cells formed haemorrhagic masses regardless of irradiation. T3359 CD133⁻ cells did not generate tumours. D456MG CD133⁻ cells (2×10^6) formed small tumours in two out of five brains. **d**, Representative photomicrographs of tumours from **c** (stained with haematoxylin and eosin). **e**, Immunofluorescent staining of frozen sections of the tumours generated by untreated and irradiated CD133⁺ cells in **c**. Cells were stained for the GFAP astrocyte marker (green) and the O4 oligodendrocyte marker (red).

xenograft or patient specimen treated with a clinically relevant IR dose (2 Gy) formed tumours with similar potency to untreated CD133⁺ cancer cells, although treatment of CD133⁺ cells with 5 Gy of IR reduced tumorigenicity (Supplementary Table 2). 2-Gy-irradiated CD133⁺ tumour cells derived from human glioblastoma biopsy specimens or xenografts formed tumours with similar latencies to non-irradiated CD133⁺ tumour cells (Fig. 2c–e and Supplementary Fig. S13a). In addition, irradiated CD133⁺ cells retained multilineage differentiation potential (Supplementary Fig. S7b) and formed tumours with heterogeneous tumour cell populations expressing markers of different lineages *in vivo* (Fig. 2d, e). Viable CD133⁺ cells from *in vivo* irradiated xenografts or patient specimens transiently grown in mice formed secondary tumours with similar latencies to CD133⁺ tumour cells from matched non-irradiated control tumours (Supplementary Fig. S13b). By contrast, the same dose of irradiation abolished the tumorigenic capacity of

D456MG CD133[−] tumour cells that showed weak basal tumour formation (Supplementary Fig. S13c). Thus, CD133⁺ glioma cell subpopulations are enriched with cancer stem cells resistant to radiation *in vitro* and *in vivo* in comparison to matched CD133[−] glioma cells.

Although IR damages tumour cells through several mechanisms, IR kills cancer cells primarily through DNA damage. Thus, DNA damage checkpoint responses play essential roles in cellular radiosensitivity^{14–20}. To determine the role of DNA damage checkpoint responses in glioma cancer stem cell radioresistance, we compared early DNA damage checkpoint responses in CD133⁺ and CD133[−] glioma tumour cell subpopulations. In both CD133⁺ and CD133[−] cells derived from human glioma xenografts, DNA damage induced by IR or a radiomimetic, neocarzinostatin, potentially initiated activating phosphorylation of the ataxia-telangiectasia-mutated (ATM), Rad17, Chk1 and Chk2 checkpoint proteins (Fig. 4a and Supplementary Fig. S14a). However, activating phosphorylation of these checkpoint proteins was significantly higher in CD133⁺ cells than in CD133[−] cells (Fig. 4a and Supplementary Fig. S14a), indicating that CD133⁺ cells show greater checkpoint activation in response to DNA damage. Similarly, CD133⁺ cells isolated from several human glioblastoma specimens showed a marked increase in activation of several checkpoint proteins in response to DNA damage relative to CD133[−] cells (Fig. 4b, c, and Supplementary Fig. S14b). Rad17, a

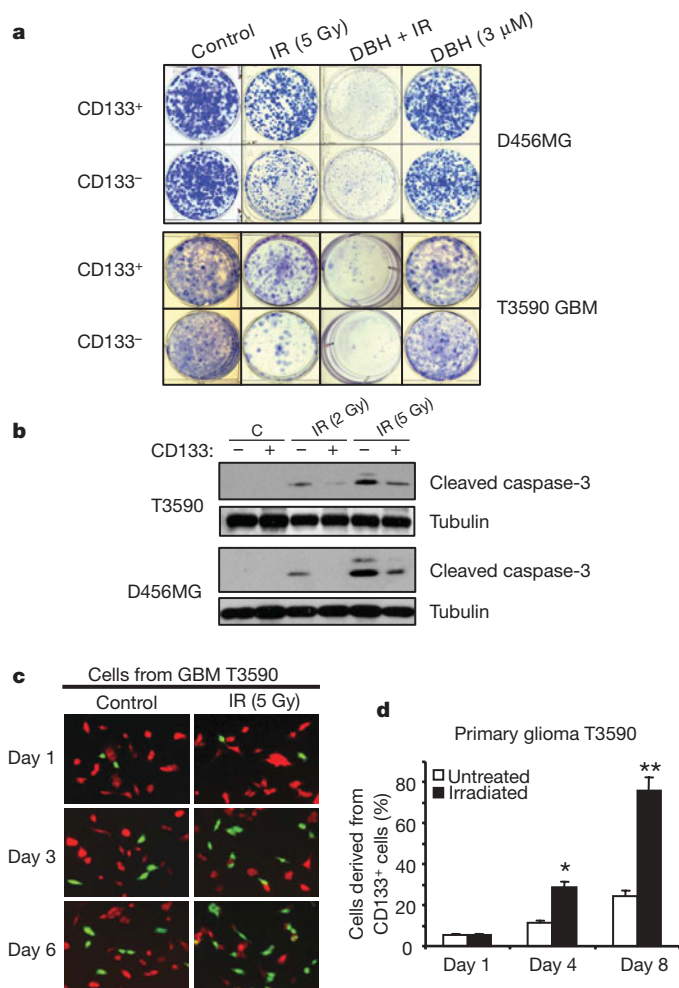


Figure 3 | CD133⁺ tumour cells show radioresistance and lower sensitivity to radiation-induced apoptosis than CD133[−] tumour cells dependent on checkpoint kinase activity. **a**, CD133⁺ and CD133[−] tumour cells were untreated or treated with 5 Gy of IR, a Chk1/Chk2 low molecular weight inhibitor (debromohymenialdisine, DBH, 3 μM), or a combination of both. Representative images of colony formation are shown. **b**, CD133⁺ and CD133[−] cells derived from primary glioblastoma T3590 or D456MG xenografts were untreated or irradiated with 2 Gy or 5 Gy of IR. Whole-cell lysates were collected after 24 h and immunoblotted for cleaved caspase-3, an indicator of cell apoptosis. **c**, CD133⁺ and CD133[−] cells derived from primary glioblastoma T3590 were labelled separately with CFSE (green) and CMTRX (red) fluorescent dyes, mixed in defined ratios (5% CD133⁺), and visualized by fluorescent microscopy at the indicated time points. **d**, Mean ± s.d. results from **c** ($n = 100$ cells in three trials; * $P < 0.002$; ** $P < 0.001$).

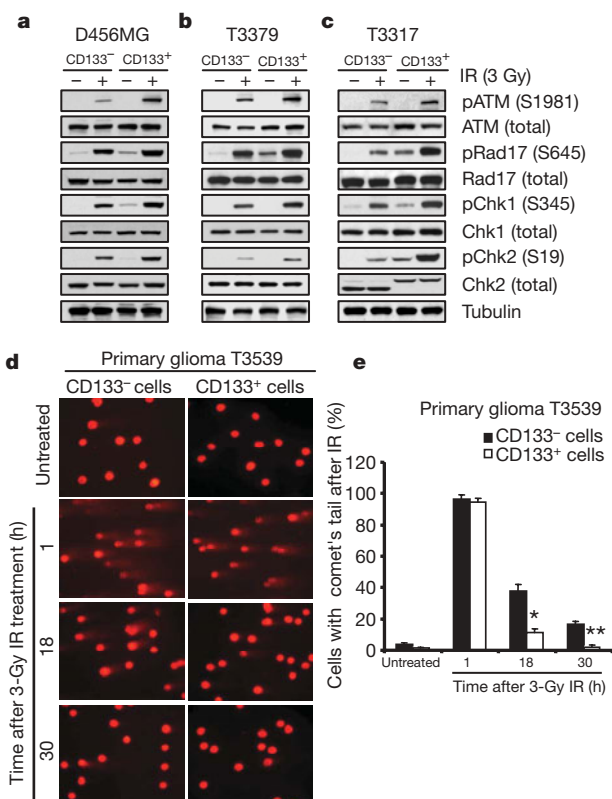


Figure 4 | CD133⁺ glioma cells preferentially activate the DNA damage checkpoint and repair IR-induced DNA damage more efficiently than CD133[−] cells. **a–c**, The activation state of the checkpoint response in matched CD133⁺ and CD133[−] cells from glioma D456MG xenografts (**a**) and glioblastoma specimens (**b**, **c**) was assessed before treatment (−) and 1 h after 3 Gy of IR (+). Whole-cell lysates were immunoblotted for phosphorylated and total amounts of checkpoint proteins (ATM, Rad17, Chk2 and Chk1). **d**, CD133⁺ and CD133[−] cells derived from primary glioblastoma T3590 were irradiated with 3 Gy of IR. The presence of DNA damage at sequential time points after damage was assessed by single-cell gel electrophoresis assay under alkaline conditions (alkaline comet assay). **e**, Quantification of the percentages of cells with comet tails at different time points after IR in CD133⁺ and CD133[−] populations. Data are the means ± s.d. ($n = 100$ cells in three trials; * $P < 0.001$; ** $P < 0.002$).

crucial regulator of the DNA damage checkpoint¹⁹, also showed significantly greater baseline phosphorylation of a key regulatory residue (Ser 645) in CD133⁺ cells than in CD133⁻ cells in most cases, suggesting that the CD133⁺ subpopulation may be primed to respond to DNA damage. In addition, IR distinctly increased ATM kinase activity in the CD133⁺ subpopulation derived from primary glioblastoma biopsy specimens in response to IR (Supplementary Fig. S14c). These data show that CD133⁺ glioma cells can activate checkpoint responses to a greater extent than CD133⁻ cells, suggesting that the resistance of CD133⁺ cells to IR is due to preferential checkpoint activation.

The primary downstream effect of checkpoint activation is to induce cell-cycle arrest to repair damaged DNA. We compared the recovery of CD133⁺ and CD133⁻ cells derived from both xenografts and biopsy specimens in response to IR-induced DNA damage using the alkaline comet assay²¹. CD133⁺ and CD133⁻ cells from glioma xenografts or human patient biopsy specimens were equally susceptible to DNA damage to IR initially, but the percentage of cells with comet tails decreased 4–9 times more rapidly in CD133⁺ cells than in matched CD133⁻ cells (Fig. 4d, e, and Supplementary Fig. S15), indicating that CD133⁺ cells repaired the DNA damage more efficiently than CD133⁻ cells. This result was further confirmed by assessing the resolution of phosphorylated histone 2AX nuclear foci²² after IR (Supplementary Fig. S16). The ability to repair DNA damage is essential to cellular survival because maintained DNA breaks induce apoptosis or senescence^{14–16}. Thus, the CD133⁺ tumour cells would be expected preferentially to survive radiation to repopulate the tumour.

To determine the contribution of the preferential activation of DNA damage checkpoint to the increased survival of CD133⁺ cells after IR administration, we assessed the ability of an inhibitor, debromohymenialdisine (DBH), of checkpoint kinases Chk1 and Chk2 (ref. 23) to prevent CD133⁺ cell resistance to IR. Pretreatment with DBH minimally impacted the proliferation of both CD133⁺ and CD133⁻ cells, but DBH treatment showed synergy with IR to disrupt the radioresistance of CD133⁺ cells *in vitro* (Fig. 3a). Similar results were obtained in *in vivo* mouse studies (Supplementary Fig. S17). These data confirm that preferential checkpoint response in CD133⁺ cancer cells is closely associated with the cellular resistance to radiation.

Together, our results show that CD133⁺ cancer cells contribute to glioma radioresistance and tumour repopulation through preferential checkpoint response and DNA repair, and targeting of checkpoint response in CD133⁺ cancer cells can overcome glioma radioresistance *in vitro* and *in vivo*, which may provide a therapeutic advantage to reduce brain tumour recurrence. The specific molecular mechanism for the resistance of cancer stem cells to IR can be linked to the function of the DNA damage checkpoint. As the cell cycle of a normal stem cell is tightly controlled by the checkpoint to maintain genomic stability and integrity, the defective checkpoint responses associated with early cancer development^{24,25} implicate abnormal checkpoint control as a potential contributor to the transformation of normal cells into cancer stem cells. Future studies may define similar contributions of the cancer stem cell to therapeutic responses in other solid cancers with similar molecular mechanisms. Therapies targeted to the checkpoint kinases in preclinical and clinical development may provide a specific method to disrupt this resistance mechanism to improve overall tumour control with radiation treatment.

METHODS

Additional and more detailed methods are presented in the Supplementary information.

Human glioma xenografts and glioblastoma specimens. D54MG xenografts were derived from A172. D456MG xenografts were derived from a paediatric glioblastoma biopsy directly implanted into immunocompromised mice. Primary glioblastoma samples (designated T3xxx) were obtained from patients

undergoing resection in accordance with a protocol approved by the Duke University Medical Center Institutional Review Board.

Isolation of CD133⁺ and CD133⁻ tumour cells. Matched subpopulations were separated as described^{3,5} with modifications²⁶.

Alkaline comet assay. DNA damage repair was assessed by single-cell gel electrophoresis assay under alkaline conditions as described²¹.

Immunofluorescent staining. Immunofluorescent staining for neural stem cell markers (CD133, Nestin, Sox2 and Musashi) and differentiation markers (GFAP and S100 β , astrocytes; O4 and GalC (Galactocerebroside), oligodendrocytes; Map-1 and TUJ1, neurons) was done as described²⁷.

Radiation treatment. Cells or mice were irradiated at indicated doses with an X-RAD 320 irradiation system (AGFA).

***In vitro* cell mixing and repopulation.** CD133⁺ derived from glioma xenografts or primary glioblastoma specimens were labelled with CellTracker CFSE green fluorescent dye (Molecular Probes), and the CD133⁻ cells were labelled with the CellTracker Red CMTPX (Molecular Probes) separately according to the instructions, and then mixed in defined ratios. Triplicate parallel cultures were left untreated or treated with IR (5 Gy). The resulting growth patterns of each tumour cell population at indicated time point was visualized by fluorescent microscopy and analysed by FACS.

Western analysis. Immunoblotting was done as described^{18,19}.

Intracranial tumour assays. Intracranial transplantation of CD133⁺ or CD133⁻ cells into brains of nude mice was done as described^{13,26}.

Statistical analysis. The level of significance was determined by a two-tailed Student's *t*-test or analysis of variance with $\alpha = 0.05$ (GraphPad software). All quantitative data presented are the mean $\pm$ s.d. from at least three samples per data point.

Received 1 June; accepted 7 September 2006.

Published online 18 October 2006.

- Legler, J. M. *et al.* Brain and other central nervous system cancers: recent trends in incidence and mortality. *J. Natl Cancer Inst.* **91**, 1382–1390 (1999).
- Garden, A. S. *et al.* Outcome and patterns of failure following limited-volume irradiation for malignant astrocytomas. *Radiother. Oncol.* **20**, 99–110 (1991).
- Singh, S. K. *et al.* Identification of human brain tumour initiating cells. *Nature* **432**, 396–401 (2004).
- Hemmati, H. D. *et al.* Cancerous stem cells can arise from pediatric brain tumours. *Proc. Natl Acad. Sci. USA* **100**, 15178–15183 (2003).
- Singh, S. K. *et al.* Identification of a cancer stem cell in human brain tumours. *Cancer Res.* **63**, 5821–5828 (2003).
- Uchida, N. *et al.* Direct isolation of human central nervous system stem cells. *Proc. Natl Acad. Sci. USA* **97**, 14720–14725 (2000).
- Galli, R. *et al.* Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma. *Cancer Res.* **64**, 7011–7021 (2004).
- Izant, J. G., Weatherbee, J. A. & McIntosh, J. R. A microtubule-associated protein antigen unique to mitotic spindle microtubules in PtK1 cells. *J. Cell Biol.* **96**, 424–434 (1983).
- Ignatova, T. N. *et al.* Human cortical glial tumours contain neural stem-like cells expressing astroglial and neuronal markers *in vitro*. *Glia* **39**, 193–206 (2002).
- Sanai, N. *et al.* Unique astrocyte ribbon in adult human brain contains neural stem cells but lacks chain migration. *Nature* **427**, 740–744 (2004).
- Reynolds, B. A. & Weiss, S. Generation of neurons and astrocytes from isolated cells of the adult mammalian central nervous system. *Science* **255**, 1707–1710 (1992).
- Oostendorp, R. A., Audet, J., Miller, C. & Eaves, C. J. Cell division tracking and expansion of hematopoietic long-term repopulating cells. *Leukemia* **13**, 499–501 (1999).
- Marin, L. *et al.* Flow cytometric quantification of apoptosis and proliferation in mixed lymphocyte culture. *Cytometry A* **51**, 107–118 (2003).
- Abraham, R. T. Cell cycle checkpoint signaling through the ATM and ATR kinases. *Genes Dev.* **15**, 2177–2196 (2001).
- Carr, A. M. DNA structure dependent checkpoints as regulators of DNA repair. *DNA Repair* **1**, 983–994 (2002).
- Zhou, B. B. & Elledge, S. J. The DNA damage response: putting checkpoints in perspective. *Nature* **408**, 433–439 (2000).
- Kastan, M. B. & Bartek, J. Cell-cycle checkpoints and cancer. *Nature* **432**, 316–323 (2004).
- Ali, A. *et al.* Requirement of protein phosphatase 5 in DNA-damage-induced ATM activation. *Genes Dev.* **18**, 249–254 (2004).
- Bao, S. *et al.* ATR/ATM-mediated phosphorylation of human Rad17 is required for genotoxic stress responses. *Nature* **411**, 969–974 (2001).
- Sancar, A., Lindsey-Boltz, L. A., Unsal-Kacmaz, K. & Linn, S. Molecular mechanisms of mammalian DNA repair and the DNA damage checkpoints. *Annu. Rev. Biochem.* **73**, 39–85 (2004).
- Tice, R. R. *et al.* Single cell gel/comet assay: guidelines for *in vitro* and *in vivo* genetic toxicology testing. *Environ. Mol. Mutagen.* **35**, 206–221 (2000).
- Balajee, A. S. & Geard, C. R. Replication protein A and γ -H2AX foci assembly is triggered by cellular response to DNA double-strand breaks. *Exp. Cell Res.* **300**, 320–334 (2004).

23. Curman, D. *et al.* Inhibition of the G2 DNA damage checkpoint and of protein kinases Chk1 and Chk2 by the marine sponge alkaloid debromohymenialdisine. *J. Biol. Chem.* **276**, 17914–17919 (2001).
24. Gorgoulis, V. G. *et al.* Activation of the DNA damage checkpoint and genomic instability in human precancerous lesions. *Nature* **434**, 907–913 (2005).
25. Bartkova, J. *et al.* DNA damage response as a candidate anti-cancer barrier in early human tumorigenesis. *Nature* **434**, 864–870 (2005).
26. Bao, S. *et al.* Stem cell-like glioma cells promote tumor angiogenesis through vascular endothelial growth factor. *Cancer Res.* **66**, 7843–7848 (2006).
27. Lee, A. *et al.* Isolation of neural stem cells from the postnatal cerebellum. *Nature Neurosci.* **8**, 723–729 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Y. H. Sun, S. Keir, D. Satterfield, L. Ehinger and J. Faison for technical assistance; M. Cook and T. R. Dissanayake for assistance with flow cytometry; Z. Lu for assistance with fluorescent microscopy; and X.-F. Wang, H. Lin, T.P. Yao, H. Friedman and R. Wechsler-Reya for discussions.

Financial support was provided by the Childhood Brain Tumor Foundation, the Pediatric Brain Tumor Foundation of the United States, Accelerate Brain Cancer Cure, a grant from the Duke Comprehensive Cancer Center Kislak–Fields Family Fund (to J.N.R.), and grants from the NIH (to J.N.R. and to D.D.B.). J.N.R. is a Damon Runyon–Lilly Clinical Investigator supported by the Damon Runyon Cancer Research Foundation and a Sidney Kimmel Foundation for Cancer Research Scholar. A.B.H. is a Paul Brazen/American Brain Tumor Association Fellow.

Author Contributions Q.W., S.B., Y.H. and Q.S. did the experimental work. R.E.M. performed pathological analysis and assisted in human tumour specimen acquisition. S.B. and J.N.R. wrote the paper and designed the experiments. A.B.H., M.W.D. and D.D.B. provided intellectual input and helped with experimental design.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.N.R. (rich0001@mc.duke.edu).

Bone morphogenetic proteins inhibit the tumorigenic potential of human brain tumour-initiating cells

S. G. M. Piccirillo^{1,2}, B. A. Reynolds³, N. Zanetti², G. Lamorte², E. Binda⁴, G. Broggi⁵, H. Brem⁶, A. Olivi⁶, F. Dimeco^{5,6} & A. L. Vescovi^{1,2,4}

Transformed, oncogenic precursors, possessing both defining neural-stem-cell properties and the ability to initiate intracerebral tumours, have been identified in human brain cancers¹. Here we report that bone morphogenetic proteins (BMPs), amongst which BMP4 elicits the strongest effect, trigger a significant reduction in the stem-like, tumour-initiating precursors of human glioblastomas (GBMs). Transient *in vitro* exposure to BMP4 abolishes the capacity of transplanted GBM cells to establish intracerebral GBMs. Most importantly, *in vivo* delivery of BMP4 effectively blocks the tumour growth and associated mortality that occur in 100% of mice after intracerebral grafting of human GBM cells. We demonstrate that BMPs activate their cognate receptors (BMPRs) and trigger the Smad signalling cascade in cells isolated from human glioblastomas (GBMs). This is followed by a reduction in proliferation, and increased expression of markers of neural

differentiation, with no effect on cell viability. The concomitant reduction in clonogenic ability, in the size of the CD133⁺ population and in the growth kinetics of GBM cells indicates that BMP4 reduces the tumour-initiating cell pool of GBMs. These findings show that the BMP–BMPR signalling system—which controls the activity of normal brain stem cells^{2,3}—may also act as a key inhibitory regulator of tumour-initiating, stem-like cells from GBMs and the results also identify BMP4 as a novel, non-cytotoxic therapeutic effector, which may be used to prevent growth and recurrence of GBMs in humans.

Within the past decade accumulating evidence from a number of biological systems, such as the blood⁴, breast⁵ and brain^{6–10}, has indicated that transformed, stem-like precursors may represent the cells that drive tumour growth¹¹. If tumour-initiating precursors with stem-like properties (TICs) are major culprits in tumour initiation

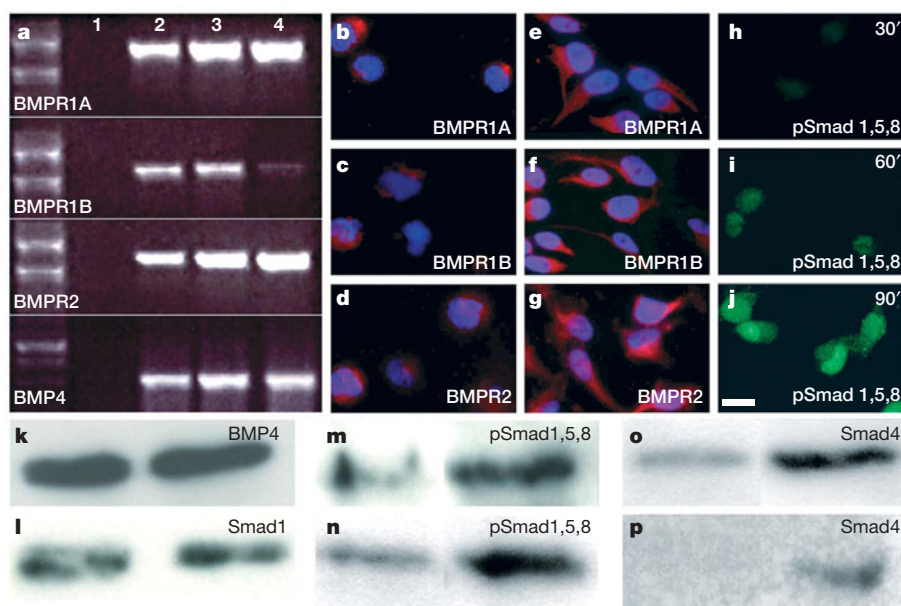


Figure 1 | Expression and activation of BMP receptors in human GBM cells. **a**, Messenger RNA transcripts for BMP receptors and BMP4. Lane 1, negative control; lane 2, acutely dissociated GBM CD133⁺ cells; lane 3, briefly cultured CD133⁺ cells; lane 4, MCF7 cells (positive control). **b–g**, BMPR immunoreactivity in acutely isolated (**b–d**) or briefly cultured (**e–g**) CD133⁺ cells. **h–j**, Phosphorylation of Smad proteins (using anti-phospho(p)Smad1,5,8) after exposure of briefly cultured cells to BMP4 for 30 min (**h**), 60 min (**i**) or 90 min (**j**) (Supplementary Fig. 2b; Supplementary

Table a). **b–j**, Scale bar, 15 μ m, as in **j**. **k–p**, Western blotting. **k**, BMP4 in acutely dissociated (left) or briefly cultured CD133⁺ cells (right). **l**, Unchanged Smad1 levels in BMP4-treated (right), briefly cultured cells (left, untreated control). **m, n**, Increased Smad1,5,8 phosphorylation by BMP4 (left, untreated control; right, BMP4-treated) in acutely dissociated (**m**) or briefly cultured (**n**) cells. **o, p**, Increased Smad4 expression in briefly cultured (**o**) or acutely dissociated (**p**) cells by BMP4 (untreated control, left; BMP4-treated, right).

¹Department of Biotechnology and Biosciences, University of Milan–Bicocca, 20126 Milan, Italy. ²Unit of Cancer Stem Cell Biology, StemGen Spa, University of Milan–Bicocca, 20126 Milan, Italy. ³Laboratory for Neural Stem Cell Biology, Queensland Brain Institute, University of Queensland, 4072 Brisbane, Queensland, Australia. ⁴Stem Cell Research Institute, H. S. Raffaele, 20132 Milan, Italy. ⁵National Neurological Institute ‘C. Besta’, 20133 Milan, Italy. ⁶Department of Neurosurgery, John Hopkins University, Baltimore, 21287 Maryland, USA.

and development, then conventional approaches targeting the overall population of neoplastic cells may well spare TICs owing to their idiosyncratic properties¹².

Similar to normal neural stem cells, brain TICs (BTICs) represent a small fraction of GBM cells⁸, belong to a CD133⁺ pool⁹, display self-renewal *in vitro*, generate a large number of progeny and are multipotent^{8,13,14}. BTICs can establish GBMs at the clonal level and perpetuate across serial transplantation^{8,9}.

Bone morphogenetic proteins (BMPs) elicit a plethora of actions^{15–19} and have an instructive role in the adult brain stem cell niche, favouring the acquisition of an astroglial fate^{2,3}. Given this and the involvement of BTICs in brain tumours, we examined the putative effects of BMPs on human GBMs.

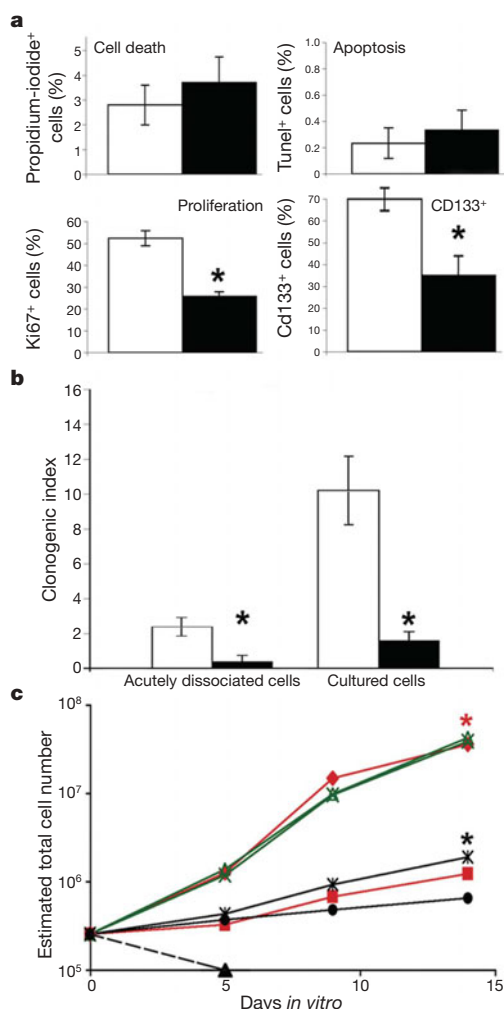


Figure 2 | BMP4 depletes the GBM BTIC population *in vitro*. **a**, BMP4 doesn't alter cell death or apoptosis (Tunel, terminal-deoxynucleotidyl-transferase-mediated dUTP nick end-labelling), but reduces proliferation (Ki67 immunoreactivity) and the CD133⁺ BTIC population⁹ in GBM cultures. Untreated control, empty column; BMP4-treated, solid column; mean $\pm$ s.e.m., $n = 3$; * $P < 0.005$. **b**, BMP4 reduces the clonogenic index in acutely dissociated and briefly cultured GBM cells. Untreated control, empty column; BMP4-treated, solid column; mean $\pm$ s.e.m., $n = 3$; * $P < 0.005$. **c**, Acutely dissociated GBM cells could not be expanded in the presence of BMP4 (black dotted line, triangles). After brief expansion with mitogens alone, acutely dissociated cells received BMP4, resulting in the reduction of the growth kinetics (red squares), similar to human fetal neural stem cells (black stars, untreated control; black circles, BMP4-treated) but unlike U87 human glioma lines (green triangles, untreated control; green stars, BMP4). Mean $\pm$ s.e.m., $n = 3$; * $P < 0.005$ for red asterisk and * $P < 0.01$ for black asterisk.

The *in vitro* findings below describe data from GBM cells^{8,9} sorted for CD133⁺ using fluorescence activated cell sorting, and either cultured for 48 h (ref. 9), and thus referred to as 'briefly cultured cells', or used soon after sorting from GBM tissue (acutely dissociated). Equivalent results were obtained with additional samples, including unsorted, briefly cultured or acutely dissociated cells from human GBMs^{8,9} (for detailed sample specifics and complete data, see Supplementary Tables a–i and respective legends). First, we demonstrated that BMP4 and its receptor (BMPR) transcripts and proteins are expressed in GBM cells, particularly in the CD133⁺ putative BTIC population⁹. Transcripts for BMPR subtypes 1A, 1B and 2 and BMPs were found in acutely dissociated and briefly cultured GBM cells (Fig. 1a; Supplementary Figs 1 and 10a). The cognate receptor (Fig. 1b–g) and BMP4 proteins (Fig. 1k) were found in both CD133⁺ and CD133[−] fractions (not shown). BMPRs were known to be functional because Smad 1–5–8 phosphorylation was observed (Fig. 1h–j and Supplementary Table a) following addition of BMP4 (100 ng ml^{−1}; Supplementary Fig. 2a); the p38 mitogen-activated protein kinase (MAPK) pathway was not engaged (Supplementary Fig. 2b).

We next investigated the effect of BMP4 on GBM cells. Unlike other cell systems²⁰, BMP4 didn't induce cell death or apoptosis, but reduced proliferation in response to mitogens (Fig. 2a; Supplementary Tables b–d). BMP2, 5, 6, 7 and 8b, but not BMP1, 3 or 3b, yielded similar effects (Supplementary Fig. 3). This was confirmed by the concomitant increase in the number of GBM cells in G0/G1 phase and a decrease of those in S phase in response to BMP4 (Supplementary Fig. 4b and Supplementary Table e). Transforming growth factor (TGF) β s, found in GBMs²¹ and known to engage overlapping signalling pathways with BMP4^{22,23}, elicited no effects (Supplementary Fig. 3), underlining the specificity of the BMP4 action on GBM cells.

Two assays—one determining the percentage of clone-forming neural precursors (the clonogenic index)²⁴ and a second assessing the expansion of the neural-stem-cell pool by growth kinetics analysis²⁵—confirmed that the cytostatic effect of BMP4 impinged on the BTIC population in GBM cells. BMP4 produced a greater than 70% reduction in the clonogenic index in GBM cells *in vitro* (Fig. 2b). Furthermore, exposure of GBM cells to BMP4 soon after isolation from the primary tumour specimen prevented their expansion in culture (Fig. 2c). Addition of BMP4 to GBM cells that were already expanding *in vitro* greatly decreased their growth rate (Fig. 2c; Supplementary Fig. 4a).

Because BMP4 reduces the growth of BTICs (Fig. 2), we investigated its effects on the CD133⁺ subpopulation, known to be enriched in BTICs⁹. BMP4 decreased the size of the CD133⁺ pool by approximately 50%, both in acutely dissociated and briefly cultured GBM cells (Fig. 2a; Supplementary Fig. 5 and Supplementary Table g). These data indicate that BMP4 targets the BTIC population.

BMP4 induced overt morphological changes in GBM cells *in vitro*. Relative to cells grown with mitogens alone, BMP4-treated cells adopted a more differentiated morphology (flat, phase-dark with elaborated processes; Fig. 3a, b; Supplementary Fig. 4c). We also observed an increase in the expression of astroglial markers (Glial fibrillary acidic protein (GFAP) immunoreactivity), together with a similar but reduced increase in neuronal (β III-tubulin and MAP5) and oligodendroglial (Galactocerebroside C (GalC) immunoreactivity) antigens (Fig. 3c–h; Supplementary Fig. 4c).

The aberrant co-expression of neuronal and glial antigens in undifferentiated BTICs makes it impractical to quantify the pro-differentiation effects of BMP4 by direct counting of immunoreactive cells (Fig. 3c, f; Supplementary Fig. 6). Therefore, we used cytofluorimetry to measure fluorescence signal intensity. Relative to control cells, BMP4-treated cultures exhibited a greater than twofold increase in GFAP immunoreactivity (Fig. 3i; Supplementary Table h) and a consistent, yet not marked, increase in β III-tubulin (Fig. 3j) and GalC immunoreactivity (Fig. 3k; Supplementary Table h).

Altogether, the results above suggested that BMP4 produces a pro-differentiation action on GBM cells (which, predominantly, acquire

an astroglial-like fate) and depletes the pool of tumorigenic BTICs. From this we inferred that the *in vitro* reduction in the stem-cell-like BTIC would correspond to a similar decline in the ability of BMP4-treated cells to form tumours *in vivo*. To test this, both acutely dissociated and briefly cultured CD133⁺ GBM cells were exposed to BMP4 for 48 h, before unilateral, intrastriatal injection (3×10^5 viable cells) into immunodeficient mice. Tumour formation and expansion were compared with those of control animals receiving GBM cells maintained under identical conditions without BMP4. All animals receiving untreated GBM cells developed large tumour masses (Fig. 4a; Supplementary Fig. 7) that showed characteristic glioblastoma features, including nuclear atypia, expression of aberrant glia, extensive neovascularization and high mitotic activity⁸ (Supplementary Fig. 7), and invaded the lateral, third and fourth ventricles (Supplementary Figs 7, 8). Conversely, BMP4-treated cells did not form invasive tumours, but small, delimited lesions, which were confined to the injection site, had a low mitotic index and showed no ventricular invasion (Fig. 4b; Supplementary Fig. 7). Three to four months post-injection all control animals died, whereas nearly all mice receiving BMP4 pre-treated cells survived (Fig. 4i). Identical results were observed when the residual CD133⁺ fraction from BMP4-treated cells was re-purified by FACS and its tumorigenicity compared with an equal number (3×10^5 CD133⁺ cells per animal) of CD133⁺ cells from matched control animals (Supplementary Fig. 7m, n). Also, we could always re-culture CD133⁺ BTICs (average clonogenic frequency: $9.0 \pm 1.3\%$ ($n = 2$), 90 days post-transplant) from the brain of mice receiving acutely dissociated control GBM cells. When re-transplanted these cells gave rise to large secondary brain tumours (see Supplementary Fig. 7). Conversely, this was never possible when animals received BMP4-pretreated GBM cells in the primary transplant, nor was it possible by direct injection of 3×10^5 cells, which were acutely dissociated from the primary tumours from these same mice. Thus, transient exposure

to BMP4 depletes the BTIC population and produces a significant decrease in the *in vivo* tumour-initiating ability of GBM cells.

Next, we sought to establish if delivery of BMP4 *in vivo* might prevent intracerebral tumour establishment and growth. Transplantation of acutely dissociated or briefly cultured GBM cells was accompanied by injection of vehicle(control)- or BMP4-saturated polyacrylic beads (releasing BMP4 for 1 week; unpublished data), at the site of cells' engraftment, either at the same time as cell implantation (co-treatment) or ten days later (post-treatment). In all experiments, animals receiving control beads developed large, malignant tumours (Fig. 4c, e, g) and died (Fig. 4i), whereas mice implanted with BMP4-releasing beads displayed small, confined lesions (Fig. 4d, f, h), surviving significantly longer (Fig. 4i). Control tumours contained pleomorphic, neoplastic elements, with reactive chromatin and malignant, infiltrating cells (Fig. 4g). Conversely, lesions in BMP4-treated animals included few neoplastic cells, but had mature elements and macrophages (Fig. 4h). The mitotic index was significantly lower in BMP4-treated animals (co-treatment: control 3.78 ± 0.17 versus BMP4 0.20 ± 0.11 ; $P < 0.01$, mean $\pm$ s.e.m., $n = 4$, two-tailed Student's *t*-test; post-treatment: control 2.88 ± 0.46 versus BMP4 0.63 ± 0.29 ; $P < 0.05$, mean $\pm$ s.e.m., $n = 4$, two-tailed Student's *t*-test). *In vivo*, immunofluorescence revealed the presence of astrocytes and nestin-positive cells, but not oligodendroglial or neuronal cells, in control and BMP4-treated tumours (Supplementary Fig. 9).

BTICs are implicated in the initiation and development of GBMs¹, the most common adult malignant brain tumour²⁶. The fact that BTICs possess some properties of normal stem cells^{12,13,27} may explain GBMs' therapy-resistance and common recurrence^{1,12,28}. Hence, discovering shared regulatory mechanisms between normal neural stem cells and BTICs may identify new targets for GBM therapy. Our data show that BMP4, a neural-stem-cell regulator, dramatically hinders the tumorigenicity of human GBM cells. The available methods used

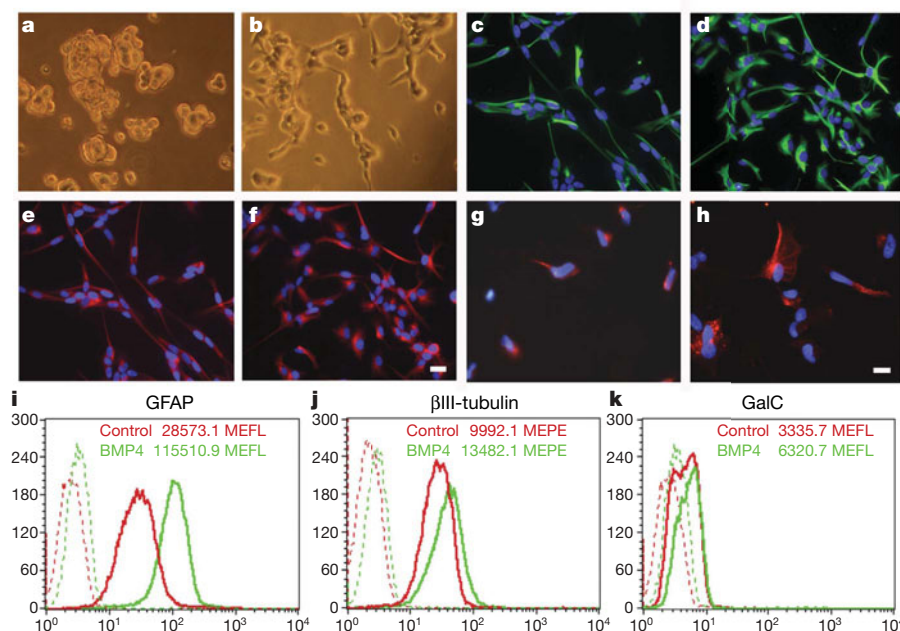


Figure 3 | BMP4 elicits a pro-differentiation effect on GBM cells.

a, b, Relative to mitogens alone (control; **a**) more differentiated (that is, adherent and phase-dark) cells (**b**) are seen on addition of BMP4 to briefly cultured cells. **c–h**, BMP4 increases GFAP immunoreactivity (**c**, untreated control; **d**, BMP4-treated), βIII-tubulin immunoreactivity (**e**, untreated control; **f**, BMP4-treated) and GalC immunoreactivity (**g**, untreated control; **h**, BMP4-treated) relative to controls. Representative micrographs of three

experiments are shown. **a–f**, Scale bar, 20 μm, as in **f, g, h**, Scale bar, 15 μm, as in **h, i–k**, Cytofluorimetric quantification in untreated control (red) versus BMP4-treated (green), briefly cultured GBM cells, revealing an increase in the molecules of equivalent fluorescein (MEFL) for GFAP immunoreactivity (**i**). A slight increase in equivalent phycoerythrin molecules (MEPE) for βIII-tubulin (**j**) or MEFL for GalC immunoreactivity (**k**), was observed (Supplementary Table h).

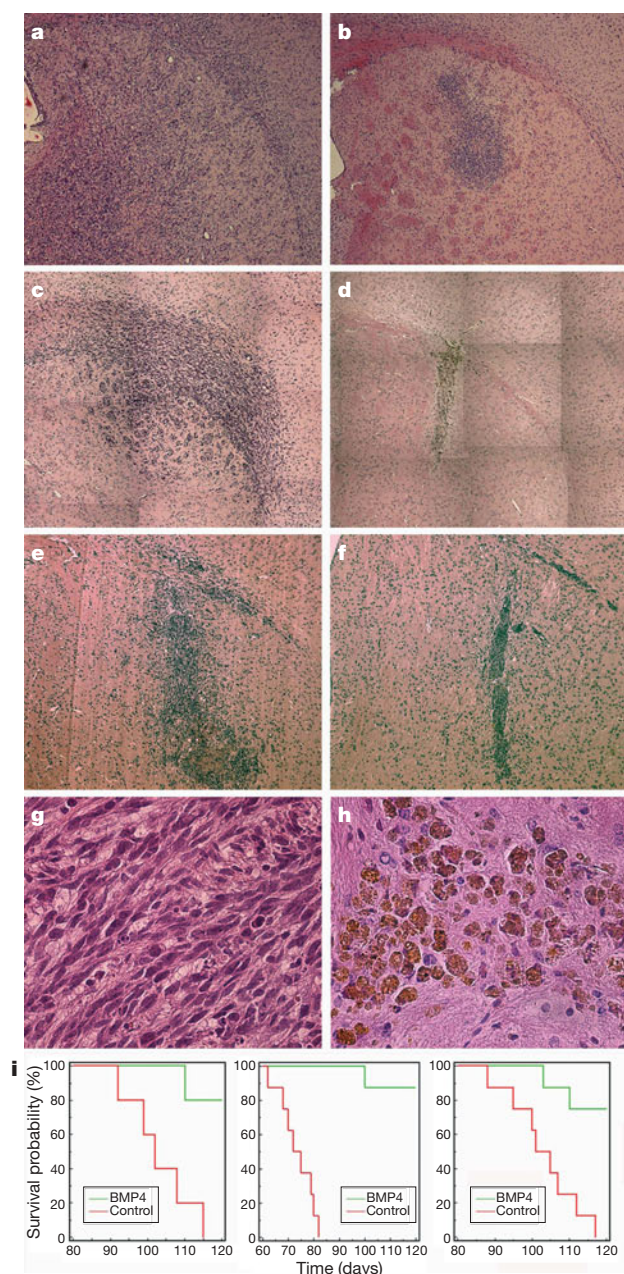


Figure 4 | BMP4 inhibits tumorigenicity of GBM cells. **a, b**, Forty-eight hour exposure of acutely dissociated GBM cells to BMP4 *in vitro*, before transplantation into the striatum of *scid/bg* mice, dramatically reduces tumour-initiating ability (**a**, untreated control; **b**, five-weeks post-injection; see Supplementary Fig. 7 for data on briefly cultured cells). **c–f**, Polyacrylic beads, pre-adsorbed with BMP4 and implanted at the cell injection site inhibit tumour growth when co-injected at the same time as cells (co-treatment; **c**, untreated control; **d**, BMP4-treated; **c, d**, briefly cultured GBM cells, four-weeks post-injection, $n = 14$) or 10 days thereafter (post-treatment; **e**, untreated control; **f**, BMP4-treated; **e, f**, acutely dissociated GBM cells, 4 weeks from cell injection, $n = 5$). The mitotic index was always significantly higher in untreated controls compared with BMP4-treated animals, in co-transplantation (3.78 ± 0.17 versus 0.20 ± 0.11 , respectively, $n = 4$; $P < 0.01$, two-tailed Student's *t*-test) and post-transplantation experiments (2.88 ± 0.46 versus 0.63 ± 0.29 , respectively, $n = 4$; $P < 0.05$, two-tailed Student's *t*-test). Tumours in untreated control mice contain pleomorphic, neoplastic and malignant, infiltrating cells (**g**). Lesions in BMP4-treated animals contained differentiated cells and macrophages (**h**). Magnification was $\times 5$ in **a–f**, and $\times 40$ in **g, h**. **i, j**, Survival of BMP4-treated animals was enhanced in pre- (left) co- (centre) and post-transplantation treatment (right) paradigms (Logrank test, $P < 0.001$, $P < 0.001$ and $P < 0.005$, respectively).

here do enable significant enrichment of GBM BTICs^{8,9}, but still yield preparations containing additional cell types. Thus, although BMP4 targets the GBM stem-like BTIC pool, it is also likely to affect these heterogeneous GBM cell populations as a whole.

Blocking endogenous BMP4 (Fig. 1 and Supplementary Fig. 10a) reduces Smad signalling and increases GBM cell growth (Supplementary Fig. 10b, c). Thus, endogenous BMP4 may operate in GBM cells, regulating the balance between proliferation and differentiation, and favouring the production of the differentiated astroglial-like cells normally found within GBMs.

Regarding the mechanisms underlying the growth inhibitory effects of BMP4 on GBM BTICs: BMP4 may reduce the frequency of BTICs by decreasing symmetric cell cycles that generate two BTICs on division; trigger differentiation of a subpopulation of BTICs; or block proliferation of BTICs and their progeny. Although not mutually exclusive, the above scenarios would all reduce the BTIC population.

We show that a critical system that regulates normal neural-stem-cell fate^{2,29} operates in the BTICs of human GBMs. It also unveils an undocumented role for BMPs in human GBM biology: their ability to promote differentiation and to deplete the pool of BTICs. The finding that *in vivo* delivered BMP4 blocks tumour development raises the potential of use in patients following surgery. These results support a new approach to GBM treatment—inducing differentiation of the tumour-initiating cells, rather than killing them. The delivery of BMP4 or BMP-mimetic drugs, combined with classical therapy, may help reduce patient lethality due to GBM.

METHODS

Samples. Data are from cells dissociated from adult human GBM specimens, obtained and classified according to the guidelines of the World Health Organization.

Primary cell culturing and propagation. Cells were isolated from GBMs as described⁸ and sorted for CD133⁺ immunoreactivity and used as such (acutely dissociated cells) or following plating in NeuroCult NS-A medium (Stem Cell Technologies) and culturing with mitogens for 48 h (briefly cultured) (see Supplementary Methods). Culture propagation, clonogenic assay and population analysis were performed as previously described^{8,9}. BMP4 pre-treatment was performed by adding 100 ng ml^{-1} of BMP4 to growth medium, 48 h before transplantation.

Immunocytochemistry. GBM cells were plated onto Matrigel-coated glass coverslips in the presence of FGF2/EGF and treated with BMP4 (100 ng ml^{-1}) for 48 h. Cells were washed, fixed in 4% paraformaldehyde and processed for immunofluorescence of neural antigens (GFAP, Dako Corporation; neuronal β -tubulin, Babco; GalC, Chemicon) as described⁸. Ki67 staining (1:1,000, Novocastra) detected proliferating cells. For the remaining immunocytochemical analysis see Supplementary Methods.

Evaluation of tumorigenicity by orthotopic injection and immunohistochemistry. Cells were orthotopically transplanted following washing and resuspension in PBS (10^8 cells per ml). Three microlitres were injected stereotactically into the right striatum of immunodeficient mice as previously described⁸. Polymer-based delivery of BMP4 used BMP4-loaded heparin acrylic beads (100 beads per animal; Sigma). Beads were incubated for 1 h at 37°C in PBS alone or containing BMP4 ($0.65 \mu\text{g } \mu\text{l}^{-1}$), rinsed three times with PBS and implanted. Haematoxylin- and eosin-staining and immunohistochemistry were performed on paraffin-embedded, 4- μm thick sections and processed as previously described^{8,30}.

Kaplan Meier survival analysis. A downward sloping plot of the cumulative chance (y-axis) of survival time was performed by MedCalc software (Mariakerke). Differences were analysed by Logrank test.

Received 6 April; accepted 12 October 2006.

- Vescovi, A. L., Galli, R. & Reynolds, B. A. Brain tumour stem cells. *Nature Rev. Cancer* **6**, 425–436 (2006).
- Lim, D. A. et al. Noggin antagonizes BMP signaling to create a niche for adult neurogenesis. *Neuron* **28**, 713–726 (2000).
- Panchision, D. M. & McKay, R. D. The control of neural stem cells by morphogenic signals. *Curr. Opin. Genet. Dev.* **12**, 478–487 (2002).
- Lapidot, T. et al. A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).
- Al-Hajj, M., Wicha, M. S., Benito-Hernandez, A., Morrison, S. J. & Clarke, M. F. Prospective identification of tumorigenic breast cancer cells. *Proc. Natl Acad. Sci. USA* **100**, 3983–3988 (2003).

6. Ignatova, T. N. *et al.* Human cortical glial tumors contain neural stem-like cells expressing astroglial and neuronal markers *in vitro*. *Glia* **39**, 193–206 (2002).
7. Hemmati, H. D. *et al.* Cancerous stem cells can arise from pediatric brain tumors. *Proc. Natl Acad. Sci. USA* **100**, 15178–15183 (2003).
8. Galli, R. *et al.* Isolation and characterization of tumorigenic, stem-like neural precursors from human glioblastoma. *Cancer Res.* **64**, 7011–7021 (2004).
9. Singh, S. K. *et al.* Identification of human brain tumour initiating cells. *Nature* **432**, 396–401 (2004).
10. Taylor, M. D. *et al.* Radial glia cells are candidate stem cells of ependymoma. *Cancer Cell* **8**, 323–335 (2005).
11. Sell, S. Stem cell origin of cancer and differentiation therapy. *Crit. Rev. Oncol. Hematol.* **51**, 1–28 (2004).
12. Reya, T., Morrison, S. J., Clarke, M. F. & Weissman, I. L. Stem cells, cancer, and cancer stem cells. *Nature* **414**, 105–111 (2001).
13. Potten, C. S. & Loeffler, M. Stem cells: attributes, cycles, spirals, pitfalls and uncertainties. Lessons for and from the crypt. *Development* **110**, 1001–1020 (1990).
14. Weiss, S. *et al.* Multipotent CNS stem cells are present in the adult mammalian spinal cord and ventricular neuroaxis. *J. Neurosci.* **16**, 7599–7609 (1996).
15. Shah, N. M., Groves, A. K. & Anderson, D. J. Alternative neural crest cell fates are instructively promoted by TGF β superfamily members. *Cell* **85**, 331–343 (1996).
16. Furuta, Y., Piston, D. W. & Hogan, B. L. Bone morphogenetic proteins (BMPs) as regulators of dorsal forebrain development. *Development* **124**, 2203–2212 (1997).
17. Li, W., Cogswell, C. A. & LoTurco, J. J. Neuronal differentiation of precursors in the neocortical ventricular zone is triggered by BMP. *J. Neurosci.* **18**, 8853–8862 (1998).
18. Iantosa, M., McPherson, C., Ho, S. & Maxwell, G. Bone morphogenetic proteins-2 and -4 attenuate apoptosis in a cerebellar primitive neuroectodermal tumor cell line. *J. Neurosci. Res.* **56**, 248–258 (1999).
19. Chen, D., Zhao, M. & Mundy, G. R. Bone morphogenetic proteins. *Growth Factors* **22**, 233–241 (2004).
20. Hallahan, A. R. *et al.* BMP-2 mediates retinoid-induced apoptosis in medulloblastoma cells through a paracrine effect. *Nature Med.* **9**, 1033–1038 (2003).
21. Kjellman, C. *et al.* Expression of TGF- β isoforms, TGF- β receptors, and SMAD molecules at different stages of human glioma. *Int. J. Cancer* **89**, 251–258 (2000).
22. Canalis, E., Economides, A. & Gazzerro, E. Bone morphogenetic proteins, their antagonists, and the skeleton. *Endocr. Rev.* **2**, 218–235 (2003).
23. Golestaneh, N. & Mishra, B. TGF- β , neuronal stem cells and glioblastoma. *Oncogene* **24**, 5722–5730 (2005).
24. Reynolds, B. A. & Weiss, S. Clonal and population analyses demonstrate that an EGF-responsive mammalian embryonic CNS precursor is a stem cell. *Dev. Biol.* **175**, 1–13 (1996).
25. Reynolds, B. A. & Rietze, R. L. Neural stem cells and neurospheres—re-evaluating the relationship. *Nature Methods* **2**, 333–336 (2005).
26. Barker, F. G. II. *et al.* Survival and functional status after resection of recurrent glioblastoma multiforme. *Neurosurgery* **42**, 709–720; discussion 720–3 (1998).
27. Dean, M., Fojo, T. & Bates, S. Tumour stem cells and drug resistance. *Nature Rev. Cancer* **5**, 275–284 (2005).
28. Pardal, R., Clarke, M. F. & Morrison, S. J. Applying the principles of stem-cell biology to cancer. *Nature Rev. Cancer* **3**, 895–902 (2003).
29. Rajan, P., Panchision, D. M., Newell, L. F. & McKay, R. D. BMPs signal alternately through a SMAD or FRAP-STAT pathway to regulate fate choice in CNS stem cells. *J. Cell Biol.* **161**, 911–921 (2003).
30. Vescovi, A. L. *et al.* Isolation and cloning of multipotential stem cells from the embryonic human CNS and establishment of transplantable human neural stem cell lines by epigenetic stimulation. *Exp. Neurol.* **156**, 71–83 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Associazione Italiana per la Ricerca sul Cancro, the Plurigenes EEC grant, funding from the Brain Tumours Funders' Collaborative, from Neurothon Onlus, and by a donation from BMW Italia. We are grateful to R. Galli, J. Laterra, C. L. Solero and F. McBlane for helpful comments and suggestions.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.L.V. (vescovi@tin.it).

LETTERS

A bacterial dynamin-like protein

Harry H. Low¹ & Jan Löwe¹

Dynamins form a superfamily of large mechano-chemical GTPases that includes the classical dynamins and dynamin-like proteins (DLPs)¹. They are found throughout the Eukarya, functioning in core cellular processes such as endocytosis and organelle division¹. Many bacteria are predicted by sequence to possess large GTPases with the same multidomain architecture that is found in DLPs². Mechanistic dissection of dynamin family members has been impeded by a lack of high-resolution structural data currently restricted to the GTPase^{3,4} and pleckstrin homology⁵ domains, and the dynamin-related human guanylate-binding protein⁶. Here we present the crystal structure of a cyanobacterial DLP in both nucleotide-free and GDP-associated conformation. The bacterial DLP shows dynamin-like qualities, such as helical self-assembly and tubulation of a lipid bilayer. *In vivo*, it localizes to the membrane in a manner reminiscent of FZL⁷, a chloroplast-specific dynamin-related protein with which it shares sequence similarity. Our results provide structural and mechanistic insight that may be relevant across the dynamin superfamily. Concurrently, we show compelling similarity between a cyanobacterial and chloroplast DLP that, given the endosymbiotic ancestry of chloroplasts⁸, questions the evolutionary origins of dynamins.

Dynamins are distinguished from classical signalling GTPases by their large size (~100 kDa)⁹, low affinity for GTP¹⁰, high basal GTP hydrolysis rate¹⁰, and self-assembly into filamentous rings¹¹ and helices¹². Minimally, all family members have a GTPase domain, middle domain and GTPase effector domain (GED)¹. GTPase activity is cooperative, with the basal rate increasing between 10- and 100-fold under conditions that promote self-assembly into oligomers^{13–15}. Dynamins promote the fission or fusion of membrane by binding to lipid through the pleckstrin homology (PH) domain in classical dynamins¹⁶ or through a functionally equivalent yet non-conserved region in DLPs located between the middle domain and GED¹. *In vitro*, helical dynamin filaments tubulate liposomes^{17,18} and coat pre-formed lipid tubes¹³.

Many Eubacteria have hypothetical genes that tentatively encode dynamin-like proteins containing a GTPase domain, middle domain and GED². Such proteins include ZP_00108538 (Genbank accession code) from the filamentous cyanobacterium *Nostoc punctiforme*, which is the subject of this study and which we term bacterial dynamin-like protein (BDLP). Purified *N. punctiforme* BDLP is a GTPase with a Michaelis constant (K_m) of 68.6 μ M and a catalytic rate constant (k_{cat}) of 0.53 min⁻¹ (Fig. 1a, b), which is in the range of the dynamin family^{1,19}. The P-loop mutant K82A reduces GTPase activity about 15-fold (Fig. 1b). Western blot analysis against *N. punctiforme* whole-cell lysate shows that this protein is produced *in vivo* (Fig. 1c).

In the presence of GMPPNP (a non-hydrolysable GTP analogue) and either *N. punctiforme* (not shown) or *Escherichia coli* whole-cell lipid extract extruded to form liposomes, BDLP strikingly self-assembles, binding to the lipid bilayer in a regular pattern (Fig. 1d–f and Supplementary Fig. 1g, h). Liposomes are tubulated in a compact helical coating with a 60 Å interfilament pitch repeat and a diameter

of ~45 nm. Broken coated tube stubs yield a detailed cross-sectional view into the helix, which appears as a cartwheel with 17-fold rotational symmetry (Fig. 1e, f). Thin radial spokes of ~80 Å merge

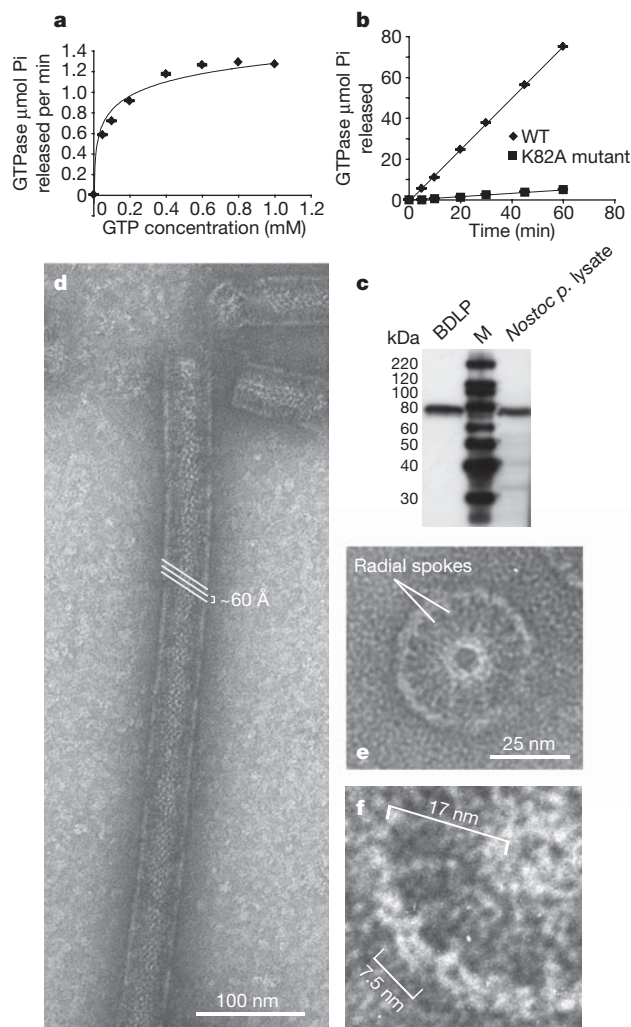


Figure 1 | BDLP is a GTPase capable of decorating and tubulating liposomes. **a**, Basal GTP hydrolysis rates for 2.5 μ M BDLP (ZP_00108538 from *N. punctiforme*) at variable GTP concentration. **b**, Time course comparing GTP hydrolysis of wild-type (WT) and K82A (Walker A) mutant BDLP. Error bars represent the s.d. from three independent experiments. **c**, Western blot analysis. Left, BDLP purifies as a single band at 79.4 kDa (including the C-terminal 6 $\times$ His tag and N-terminal Gly-Ser-His linker). Right, BDLP is expressed in vegetatively growing *N. punctiforme*. **d**, In the presence of 2 mM GMPPNP, BDLP (25 μ M) decorates and tubulates *E. coli* lipid liposomes. A 60 Å interfilament pitch repeat is deduced from diffraction analysis (not shown). **e**, **f**, Decorated tube stubs in cross-section reveal the internal organization of the helix.

¹MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK.

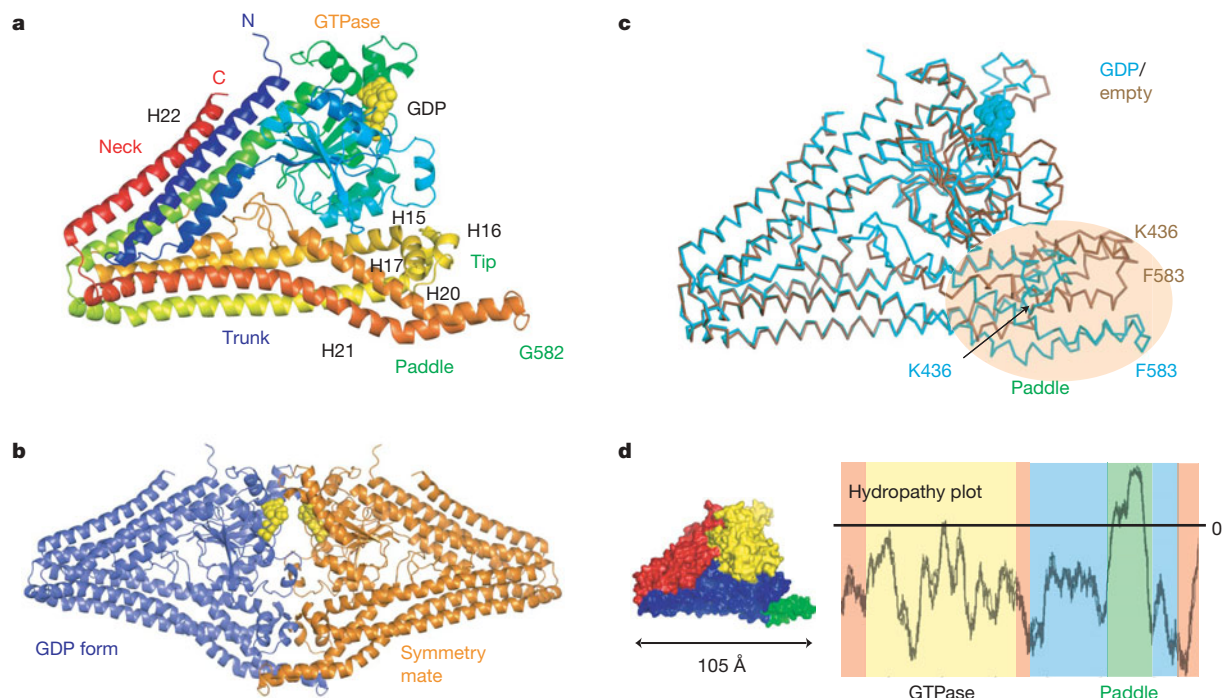


Figure 2 | Crystal structures of BDLP in GDP-associated and nucleotide-free conformations. **a**, The BDLP-GDP monomer and domain assignment. The monomer is rainbow-coloured from the N (blue) to the C (red) terminus. **b**, GDP binding promotes homodimerization, which buries the GTPase domain nucleotide-binding sites and induces tip interplay. **c**, Main chain superposition showing structural reorganization between the GDP-bound and nucleotide-free states. Note the significant conformational change at the tip. **d**, Surface plot of the GDP monomer coloured by core regions: yellow, GTPase domain; red, neck region; blue, trunk region; green, paddle region (left). Similarly colour-coded hydropathy plot (TMpred) strongly predicts that the paddle region (residues 572–606) is transmembrane (right).

into a dense outer ring forming a T-shape repeat, while presumably associating with the lipid tube on the inside. Such architecture is reminiscent of eukaryotic dynamin liposome tubulation¹⁸ and three-dimensional (3D) electron microscopy dynamin tube^{20,21} and ring²² reconstructions, which have approximate 14- and 11-fold radial symmetry, respectively. All reconstructions seem to show that the basic assembly unit is a T-shaped repeat, which is thought to comprise a dimer. Further characterization of BDLP in the presence or absence of *E. coli* lipid liposomes and different nucleotides is given in Supplementary Fig. 1 and Supplementary Information.

We did not detect any increase in GTPase activity stimulated by *N. punctiforme* or *E. coli* whole-cell lipid extract. This was unexpected given the striking lipid-induced self-assembly observed by electron microscopy. Activation may require highly specific membrane organization or other unknown cellular binding partners.

We crystallized *N. punctiforme* BDLP in a GDP-associated and a nucleotide-free state. The structures were solved to a resolution of 3.1 Å and 3.0 Å, respectively (Table 1 and Supplementary Table 1). In either form, the molecule is surprisingly compact given the electron microscopy data, and comprises a GTPase head, a four-helix neck and trunk bundle, and a tip domain (Fig. 2a and Supplementary Fig. 2e). On the basis of sequence alignments, DLPs have traditionally

been divided into conserved domains consisting of the GTPase domain, middle domain and GED; in BDLP, however, the predicted middle domain and GED do not form discrete entities and instead run in parallel, as four helix bundles, contributing to both the neck and trunk regions. The GTPase domain (residues 68–287) has a long helix–loop–helix amino terminus (1–67), which substantially buttresses the neck. The predicted GED domain at the carboxy terminus of the molecule (607–693) does not make contact with the GTPase domain and is probably involved only in oligomerization. The tip region seems highly specialized and comprises two helical layers (Fig. 2a). Helices 15, 16 and 17 (412–449), derived from the predicted middle domain, form a flexible upper layer, whereas the C terminus of helix 20 linked to the N terminus of helix 21 (572–606) forms a mobile paddle at the base of the molecule. This paddle is positioned between the predicted middle domain and GED and would be expected to mediate lipid binding. Remarkably, given the high solubility of BDLP without detergent, both paddle helices are strongly predicted to be transmembrane (Fig. 2d). Insertion of the paddle into the lipid bilayer could be used as a mechanism to promote membrane curvature. In classical eukaryotic dynamin, the additional PH domain may either replace the paddle region or act in concert by adding lipid binding specificity.

Table 1 | Refinement statistics

	Empty	GDP
Model	4 monomers per asymmetric unit; chains A–D; residues 4–501, 510–526, 540–695; 0 nucleotide; 0 water	1 monomer per asymmetric unit; chain A, residues 2–97, 103–501, 511–695; 1 GDP, 0 Mg ²⁺ ; 0 water
Diffraction data	3.0 Å, all data	3.1 Å, all data
R factor, R_{free}	0.24 (0.34), 0.28 (0.39)	0.22 (0.35), 0.27 (0.36)
B factors	65.1 Å ² , 0.5 Å ²	69.3 Å ² , 3.3 Å ²
Geometry	0.009 Å, 1.226°	0.007 Å, 1.259°
Restrained NCS	0.12 Å	NA
Protein Data Bank accession code	2J69	2J68

B factors, average and root-mean-square deviation between bonded atoms; geometry, root-mean-square deviation for bond lengths and angles from ideal values; NA, not applicable. R factor and R_{free} high-resolution bin values shown in parentheses.

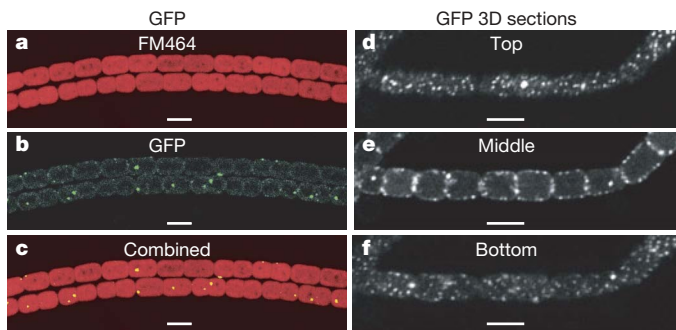


Figure 3 | BDLP shows a punctate pattern of localization in filamentous *N. punctiforme*. Scale bars, 4 μm . **a–c**, BDLP–GFP fusion forms a low number of foci at the cell periphery. **d–f**, BDLP–GFP punctate localization is restricted to the cell envelope. Shown are 0.2 μm sections taken through the top, middle and bottom of the cell filament.

Binding of GDP promotes formation of a crystallographic dimer with a substantial interface of 7.3% of the monomer surface (2,327/32,044 $\text{\AA}^2$; Fig. 2b). Homodimerization occurs between GTPase domains and occludes the nucleotide-binding site. Helix 9 reaches across the interface and Glu 245 located at the C-terminal tip of the helix is in a position to hydrogen bond with and to stabilize the *trans* nucleotide base (Supplementary Fig. 2f). Significantly, similar GTPase homodimerization across the nucleotide-binding site has been observed in the structure of the GTPase domain of another dynamin family member, human guanylate-binding protein²³ (Supplementary Fig. 2i), and may emerge to be a basic feature of the dynamin family. Homodimerization in human guanylate-binding protein generates a conformation in which Arg 48 located in the P-loop is oriented in *cis* for efficient nucleotide catalysis. Lys 79 in BDLP is found at the equivalent location and could serve a similar purpose (Supplementary Fig. 2g). Tip helices 15, 16 and 17 are heavily kinked, running almost perpendicular to the trunk axis, and are intimately interleaved with the upper paddle surface of the symmetry mate (Fig. 2b). Analytical ultracentrifugation (AUC) of BDLP in the presence of GDP shows a monomer–dimer equilibrium with a dissociation constant (K_d) of 30 μM (Supplementary Fig. 2j). The nucleotide-free monomer forms a non-crystallographic dimer also across the GTPase active site, but with a limited interface of 3.3% of the monomer surface (1,019/31,339 $\text{\AA}^2$). AUC data of BDLP in the absence of nucleotide is shown in Supplementary Fig. 2j.

Nucleotide release promotes rearrangement of the active site with Asp 98 projecting into and occluding the P-loop (Supplementary Fig. 2h). Conserved Thr 103 in the switch 1 region, which is critical for catalysis in GTPases, is orientated away from the active site. The most significant conformational change occurs, however, in the tip. Helices 15, 16 and 17 release the symmetry mate paddle and straighten, realign with the trunk axis, and occlude the upper surface of their own paddle, which has swung up to make contact, thereby providing a potential mechanism to shear the GDP dimer apart (Fig. 2c). Phe 583 located at the paddle tip moves up 21 $\text{\AA}$, whereas Lys 436 between helices 16 and 17 traverses 25 $\text{\AA}$.

The GDP dimer forms a compact diamond-shaped unit with a maximal distance of 80 $\text{\AA}$ from the base of the paddles to the top of the GTPase domains (Fig. 2b). In this conformation, BDLP is probably unable to generate the T-shaped repeats that are observed in decorated lipid tube cross-sections in the presence of GMPPNP (Fig. 1e, f). This difference is suggestive of a substantial intramolecular rearrangement in which the region between neck and trunk acts as a hinge that opens and markedly extends the molecule on GTP or lipid binding. In such a conformation, homodimerization across the GTPase domain interface would be expected, as observed in the GDP-associated BDLP crystal structure.

In vivo localization data were obtained by transforming wild-type *N. punctiforme* with a plasmid-borne translational fusion of BDLP and green fluorescent protein (GFP) driven by the wild-type promoter. Using confocal fluorescence microscopy, we observed foci localizing clearly to the cell periphery (Fig. 3a–f). The quantity of foci varied considerably, probably owing to differences in expression, ranging from just a few foci per cell (Fig. 3a–c) to an even covering of foci throughout the cell envelope (Fig. 3d–f). Preferential localization of the GFP fusion was sometimes observed at the cell septum (Supplementary Fig. 3g, h, inset) and ring-like structures could be seen (Supplementary Fig. 3h), although such phenomena were observed only rarely. Immunofluorescence microscopy using an affinity purified antibody raised against BDLP showed a similar localization pattern to the GFP fusion (Supplementary Fig. 3i–k). Foci predominantly localized to the cell membrane, although, in contrast to the GFP fusion, punctate localization was also observed in the cell interior.

Sequence alignments of BDLP with other dynamin family members show that BDLP is most closely related to the mitofusin class of DLPs, which includes *Drosophila* mitochondrial Fzo homologues^{24–27} and *Arabidopsis* chloroplast FZL⁷ (Supplementary Fig. 3l). The mitofusins have a predicted domain architecture similar to BDLP, including two transmembrane regions at the C terminus that are separated by a short linker and flanked by predicted coiled-coil domains. Alignment of BDLP and FZL (initiated at the P-loop owing to a FZL N-terminal chloroplast import sequence) appears superior to other dynamin family members, with 22% identity and 37% similarity. FZL–GFP in *Arabidopsis* chloroplasts shows a punctate localization pattern at both the chloroplast inner-envelope membrane and topologically bordering thylakoid membranes⁷. BDLP–GFP and immunofluorescence microscopy studies show a strikingly similar punctate phenotype. Given the endosymbiotic origins of chloroplasts from a cyanobacteria, the crystal structure of BDLP may represent FZL and possibly the mitofusin class of DLPs. *fzl* knockout mutants have altered chloroplast morphology, plant growth and thylakoid ultrastructure⁷, which can be rescued by FZL–GFP. BDLP may therefore have a similar role in the cyanobacterium *N. punctiforme* in determining thylakoid morphology and cell shape.

Given the presence of large GTPases with predicted dynamin-like domain organization in many members of the Eubacteria such as *E. coli* and *Bacillus subtilis*, it is likely that bacterial dynamins, or BDLPs as we term this class, are not restricted to cyanobacteria. Such observation, combined with our results, suggests a bacterial ancestry for the dynamin superfamily.

METHODS

Cloning, expression and purification. The coding sequence for BDLP from *N. punctiforme* (ZP_00108538) was cloned with a N-terminal MBP fusion, TEV cleavage site in the linker, and His tag at the C terminus. The protein was expressed in *E. coli* strain BL21(DE3). The BDLP–MBP fusion was purified with nickel-Sepharose. MBP was cleaved from BDLP and separated by gel filtration. Purified BDLP includes an additional Gly-Ser-His sequence at the N terminus and a 6 $\times$ His at the C terminus (see Supplementary Methods for more details).

Crystallization, structure determination and refinement. Initial BDLP crystallization conditions were found in a screen of 1,500 conditions set up by a high-throughput nanolitre robotic system²⁸. Native and seleno-methionine BDLP was crystallized by vapour diffusion. BDLP in the nucleotide-free state was initially solved by seleno-methionine multiwavelength anomalous diffraction, whereas BDLP–GDP was solved by molecular replacement (see Supplementary Methods for more details).

Electron microscopy and GTPase assays. *E. coli* whole-cell lipid extract was used to make liposomes. For tubulation, 2 mM GMPPNP was incubated at 37 $^{\circ}\text{C}$ with 25 μM BDLP and 2 mg ml^{-1} of liposomes. Grids were negatively stained. GTPase assays were done with a malachite-green-based kit (see Supplementary Methods for more details).

For details of western blotting, immunofluorescence microscopy, construction of BDLP–GFP reporter and GFP studies, cell culture and phylogenetic analysis, see Supplementary Methods.

Received 13 July; accepted 2 October 2006.

Published online 22 November 2006.

1. Praefcke, G. J. & McMahon, H. T. The dynamin superfamily: universal membrane tubulation and fission molecules? *Nature Rev. Mol. Cell Biol.* **5**, 133–147 (2004).
2. van der Bliek, A. M. Functional diversity in the dynamin family. *Trends Cell Biol.* **9**, 96–102 (1999).
3. Reubold, T. F. *et al.* Crystal structure of the GTPase domain of rat dynamin 1. *Proc. Natl Acad. Sci. USA* **102**, 13093–13098 (2005).
4. Niemann, H. H., Knetsch, M. L., Scherer, A., Manstein, D. J. & Kull, F. J. Crystal structure of a dynamin GTPase domain in both nucleotide-free and GDP-bound forms. *EMBO J.* **20**, 5813–5821 (2001).
5. Ferguson, K. M., Lemmon, M. A., Schlessinger, J. & Sigler, P. B. Crystal structure at 2.2 Å resolution of the pleckstrin homology domain from human dynamin. *Cell* **79**, 199–209 (1994).
6. Prakash, B., Praefcke, G. J., Renault, L., Wittinghofer, A. & Herrmann, C. Structure of human guanylate-binding protein 1 representing a unique class of GTP-binding proteins. *Nature* **403**, 567–571 (2000).
7. Gao, H., Sage, T. L. & Osteryoung, K. W. FZL, an FZO-like protein in plants, is a determinant of thylakoid and chloroplast morphology. *Proc. Natl Acad. Sci. USA* **103**, 6759–6764 (2006).
8. McFadden, G. I. Endosymbiosis and evolution of the plant cell. *Curr. Opin. Plant Biol.* **2**, 513–519 (1999).
9. Obar, R. A., Collins, C. A., Hammarback, J. A., Shpetner, H. S. & Vallee, R. B. Molecular cloning of the microtubule-associated mechanochemical enzyme dynamin reveals homology with a new family of GTP-binding proteins. *Nature* **347**, 256–261 (1990).
10. Song, B. D. & Schmid, S. L. A molecular motor or a regulator? Dynamin's in a class of its own. *Biochemistry* **42**, 1369–1376 (2003).
11. Hinshaw, J. E. & Schmid, S. L. Dynamin self-assembles into rings suggesting a mechanism for coated vesicle budding. *Nature* **374**, 190–192 (1995).
12. Carr, J. F. & Hinshaw, J. E. Dynamin assembles into spirals under physiological salt conditions upon the addition of GDP and γ -phosphate analogues. *J. Biol. Chem.* **272**, 28030–28035 (1997).
13. Stowell, M. H., Marks, B., Wigge, P. & McMahon, H. T. Nucleotide-dependent conformational changes in dynamin: evidence for a mechanochemical molecular spring. *Nature Cell Biol.* **1**, 27–32 (1999).
14. Warnock, D. E., Hinshaw, J. E. & Schmid, S. L. Dynamin self-assembly stimulates its GTPase activity. *J. Biol. Chem.* **271**, 22310–22314 (1996).
15. Tuma, P. L. & Collins, C. A. Activation of dynamin GTPase is a result of positive cooperativity. *J. Biol. Chem.* **269**, 30842–30847 (1994).
16. Salim, K. *et al.* Distinct specificity in the recognition of phosphoinositides by the pleckstrin homology domains of dynamin and Bruton's tyrosine kinase. *EMBO J.* **15**, 6241–6250 (1996).
17. Danino, D., Moon, K. H. & Hinshaw, J. E. Rapid constriction of lipid bilayers by the mechanochemical enzyme dynamin. *J. Struct. Biol.* **147**, 259–267 (2004).
18. Sweitzer, S. M. & Hinshaw, J. E. Dynamin undergoes a GTP-dependent conformational change causing vesiculation. *Cell* **93**, 1021–1029 (1998).
19. Narayanan, R., Leonard, M., Song, B. D., Schmid, S. L. & Ramaswami, M. An internal GAP domain negatively regulates presynaptic dynamin *in vivo*: a two-step model for dynamin function. *J. Cell Biol.* **169**, 117–126 (2005).
20. Zhang, P. & Hinshaw, J. E. Three-dimensional reconstruction of dynamin in the constricted state. *Nature Cell Biol.* **3**, 922–926 (2001).
21. Chen, Y. J., Zhang, P., Egelman, E. H. & Hinshaw, J. E. The stalk region of dynamin drives the constriction of dynamin tubes. *Nature Struct. Mol. Biol.* **11**, 574–575 (2004).
22. Klockow, B. *et al.* The dynamin A ring complex: molecular organization and nucleotide-dependent conformational changes. *EMBO J.* **21**, 240–250 (2002).
23. Ghosh, A., Praefcke, G. J., Renault, L., Wittinghofer, A. & Herrmann, C. How guanylate-binding proteins achieve assembly-stimulated processive cleavage of GTP to GMP. *Nature* **440**, 101–104 (2006).
24. Hales, K. G. & Fuller, M. T. Developmentally regulated mitochondrial fusion mediated by a conserved, novel, predicted GTPase. *Cell* **90**, 121–129 (1997).
25. Hermann, G. J. *et al.* Mitochondrial fusion in yeast requires the transmembrane GTPase Fzo1p. *J. Cell Biol.* **143**, 359–373 (1998).
26. Santel, A. & Fuller, M. T. Control of mitochondrial morphology by a human mitofusin. *J. Cell Sci.* **114**, 867–874 (2001).
27. Rapaport, D., Brunner, M., Neupert, W. & Westermann, B. Fzo1p is a mitochondrial outer membrane protein essential for the biogenesis of functional mitochondria in *Saccharomyces cerevisiae*. *J. Biol. Chem.* **273**, 20150–20155 (1998).
28. Stock, D., Perisic, O. & Löwe, J. Robotic nanolitre protein crystallisation at the MRC Laboratory of Molecular Biology. *Prog. Biophys. Mol. Biol.* **88**, 311–327 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We acknowledge support on beamlines ID29, ID14eh4 and ID23eh1 at the ESRF. We thank J. Meeks and K. Hagan for supplying bacterial strains and *N. punctiforme* plasmids including pSCR202; S. Reichelt for support with confocal microscopy; and J. Butler for performing the analytical ultracentrifugation experiments. We thank the MRC for a PhD student fellowship to H.H.L.

Author Information Atomic coordinates produced in this study have been deposited in the Protein Data Bank under accession codes 2J69 (BDLP) and 2J68 (BDLP-GDP). Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.L. (jyl@mrc-lmb.cam.ac.uk).

LETTERS

Klotho converts canonical FGF receptor into a specific receptor for FGF23

Itaru Urakawa¹, Yuji Yamazaki¹, Takashi Shimada¹, Kousuke Iijima¹, Hisashi Hasegawa¹, Katsuya Okawa¹, Toshiro Fujita², Seiji Fukumoto² & Takeyoshi Yamashita¹

FGF23 is a unique member of the fibroblast growth factor (FGF) family because it acts as a hormone that derives from bone and regulates kidney functions, whereas most other family members are thought to regulate various cell functions at a local level^{1–5}. The renotropic activity of circulating FGF23 indicates the possible presence of an FGF23-specific receptor in the kidney⁶. Here we show that a previously undescribed receptor conversion by Klotho, a senescence-related molecule⁷, generates the FGF23 receptor. Using a renal homogenate, we found that Klotho binds to FGF23. Forced expression of Klotho enabled the high-affinity binding of FGF23 to the cell surface and restored the ability of a renal cell line to respond to FGF23 treatment. Moreover, FGF23 incompetence was induced by injecting wild-type mice with an anti-Klotho monoclonal antibody. Thus, Klotho is essential for endogenous FGF23 function. Because Klotho alone seemed to be incapable of intracellular signalling, we searched for other components of the FGF23 receptor and found FGFR1(IIIc), which was directly converted by Klotho into the FGF23 receptor. Thus, the concerted action of Klotho and FGFR1(IIIc) reconstitutes the FGF23 receptor. These findings provide insights into the diversity and specificity of interactions between FGF and FGF receptors.

FGF23 has an FGF-like domain in its amino-terminal portion and has been classified as the newest member of the FGF family^{1,3}. Circulating FGF23 comes mainly from cells derived from the osteoblast lineage⁸ and helps to maintain homeostasis in phosphate and vitamin D metabolism^{9,10} by regulating the sodium phosphate co-transporter and the key vitamin D-metabolizing enzymes CYP27B1 and CYP24A1 in the kidney⁶. Because FGF23 is a systemic humoral factor, its renotropic function indicates that the kidney may bear a unique receptor specific for FGF23. Despite previous attempts with the use of renal cell lines (Supplementary Discussion), the molecular mechanism responsible for the renotropic action of FGF23 remains unclear.

To address the molecular events accounting for the renal specificity of FGF23 action, we sought to explain the early response to FGF23 *in vivo*. Using a complementary DNA array, we discovered that intravenous injection of mice with FGF23 protein significantly upregulated the expression of the gene *early growth-responsive 1* (*Egr-1*) in the kidney within 1 h. This induction of *Egr-1* was confirmed by northern blot analysis (Fig. 1a). Phosphorylation of extracellular signal-regulated kinase (ERK) also occurred as early as within 10 min, which indicates that the activation of ERK may, at least partly, mediate the induction of *Egr-1* expression (Fig. 1b). Although most cells in the body are known to be capable of ERK phosphorylation and *Egr-1* induction^{11,12}, these responses were not observed after FGF23 treatment in liver, heart, lung, spleen, stomach, small intestine, thymus, adrenal gland, brain or skeletal muscle

(Fig. 1c). Thus, FGF23 treatment activates ERK and induces *Egr-1* expression in the kidney, probably through an unknown FGF23 receptor system that is specifically present in the kidney but not in culture systems *in vitro* (Supplementary Table 1).

We then sought to identify a molecule in murine renal homogenates that directly binds to FGF23. Among several proteins adsorbed by the FGF23-Sepharose, matrix-assisted laser desorption/ionization–time-of-flight (MALDI–TOF) mass spectrometry analysis revealed that the major protein was Klotho (Fig. 2a). Klotho is a type I membrane protein that has a large extracellular domain consisting of about 980 amino acid residues and a very short 11-residue cytoplasmic region^{7,13,14}. Klotho is predominantly expressed in the kidney but not expressed in renal cell lines that we previously employed. We therefore investigated the functional relationships between Klotho and FGF23 further.

The binding assay revealed that high-affinity and low-affinity binding sites for FGF23 emerged by expressing Klotho in Chinese hamster ovary or HEK-293 cells (Fig. 2b, and Supplementary Fig. 2a, b, e). Such specific binding sites of FGF23 were not observed in opossum kidney proximal tubule (OK) cells even in the presence of heparin (Supplementary Fig. 2c, d), whereas FGF23 was reported to bind to OK cells in a glycosaminoglycan-assisted manner¹⁵ or by an unknown mechanism¹⁶ (Supplementary Discussion).

Cells transiently expressing Klotho *in vitro* became responsive to FGF23, because FGF23 increased the phosphorylation of ERK protein, *Egr-1* messenger RNA and protein abundance in a dose-dependent manner (Fig. 2c, d, and Supplementary Fig. 3). Moreover, we show by luciferase reporter assays that FGF23 could activate the promoter derived from the murine *Egr-1* gene (–1023 to +1)¹⁷ in a dose-dependent manner only in the presence of Klotho (Fig. 2e). Similar doses of basic FGF (bFGF) stimulated the phosphorylation of ERK, the expression of *Egr-1*, and *Egr-1* promoter activity independently of Klotho, and the presence of Klotho was somewhat suppressive towards bFGF signalling (Fig. 2c, d, f, and Supplementary Fig. 3).

Further indicative of the vital function of Klotho in FGF23 signalling is the observation that when rats were treated with FGF23, not only was the rapid induction of *Egr-1* just after treatment observed in the kidney, it was also detected in the pituitary and parathyroid glands, which also express Klotho abundantly^{7,13} (Fig. 1d). FGF23 signalling *in vivo* therefore also seems to depend on the expression of Klotho.

The functional relationship between Klotho and FGF23 is also suggested by phenotypic similarities between Klotho-deficient (*kl/kl*)⁷ and FGF23-null mice^{18,19}. Klotho-deficient (*kl/kl*) mice have previously been reported to show various senescence-like phenotypes such as a short lifespan, infertility, atrophy of lymphopoietic and reproductive organs, decreased bone mineral density, and

¹Pharmaceutical Research Laboratories, Kirin Brewery Co., Ltd, 3 Miyahara, Takasaki, Gunma 370-1295, Japan. ²Division of Nephrology and Endocrinology, Department of Internal Medicine, University of Tokyo Hospital, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan.

ectopic calcification; all of these characteristics are also observed in FGF23-null mice. Moreover, it was shown that FGF23 deficiency increases renal phosphate reabsorption and 1,25-dihydroxyvitamin D ($1,25(\text{OH})_2\text{D}$) synthesis that leads to elevated serum concentrations of phosphate, $1,25(\text{OH})_2\text{D}$ and calcium^{18,19}. These altered serum conditions were also observed in Klotho-deficient (*kl/kl*) mice^{20,21} (Supplementary Table 2), and restriction of vitamin D improved the Klotho-deficient phenotype, indicating that the abnormal mineral metabolism might be the primary cause of the ageing phenotype²⁰. As with other systemic endocrine hormones, FGF23 has a negative feedback loop to control its serum concentration. Once its action was impaired as described in patients with renal failure²², FGF23 production and serum concentrations increased. We found a slight increase in serum FGF23 concentrations in mice with decreased Klotho gene expression (*kl/+* and *Fgf23(+/-)(kl/+)*; Supplementary Tables 3 and 4). Furthermore, the mean serum concentration of FGF23 in Klotho-deficient (*kl/kl*) mice was $231 \pm 40 \text{ ng ml}^{-1}$, which is 2,000-fold that in wild-type mice ($113 \pm 11 \text{ pg ml}^{-1}$). The accumulated FGF23 was functional even when diluted 1,000-fold (Supplementary Fig. 4), whereas Klotho mice were unresponsive to the injection of even $150 \mu\text{g kg}^{-1}$ recombinant FGF23 (data not shown).

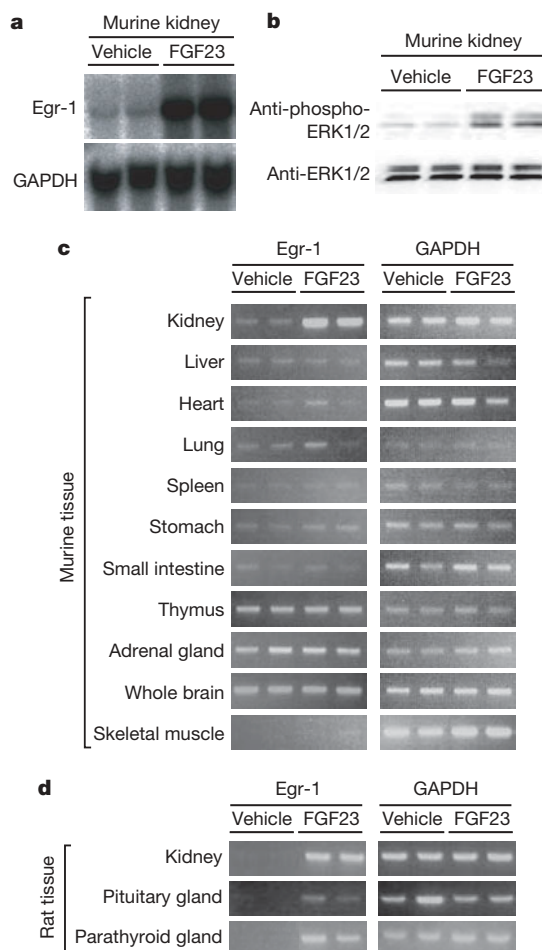


Figure 1 | *Egr-1* expression and ERK phosphorylation induced by FGF23 *in vivo*. FGF23 ($150 \mu\text{g kg}^{-1}$) or vehicle was injected into mice or rats intravenously. **a**, Murine renal *Egr-1* mRNA abundance 1 h after injection was determined by northern blotting ($n = 2$ in each case). **b**, ERK phosphorylation in the murine kidney 10 min after injection was analysed by immunoblotting ($n = 2$ in each case). **c**, *Egr-1* mRNA abundance in various murine organs 30 min after injection was determined by RT-PCR ($n = 2$ in each case). **d**, Increased *Egr-1* expression in the kidney, pituitary glands and parathyroid glands in rats 30 min after injection was determined by RT-PCR ($n = 2$ in each case). GAPDH, glyceraldehyde-3-phosphate dehydrogenase.

To clarify the importance of the interactions between Klotho and FGF23 further, we obtained a rat monoclonal antibody (70E2) against the extracellular portion of the murine Klotho protein that specifically recognizes Klotho-expressing HEK-293 cells (Fig. 3a). It also blocked the FGF23-induced increase in *Egr-1* promoter activity in cells expressing Klotho (Fig. 3b) without affecting the bFGF-dependent activation of *Egr-1* expression (data not shown). The ability of this antibody to abrogate the interaction between FGF23 and Klotho *in vivo* was then assessed by intravenously injecting wild-type mice with it. Serum concentrations of $1,25(\text{OH})_2\text{D}$ increased 9 h

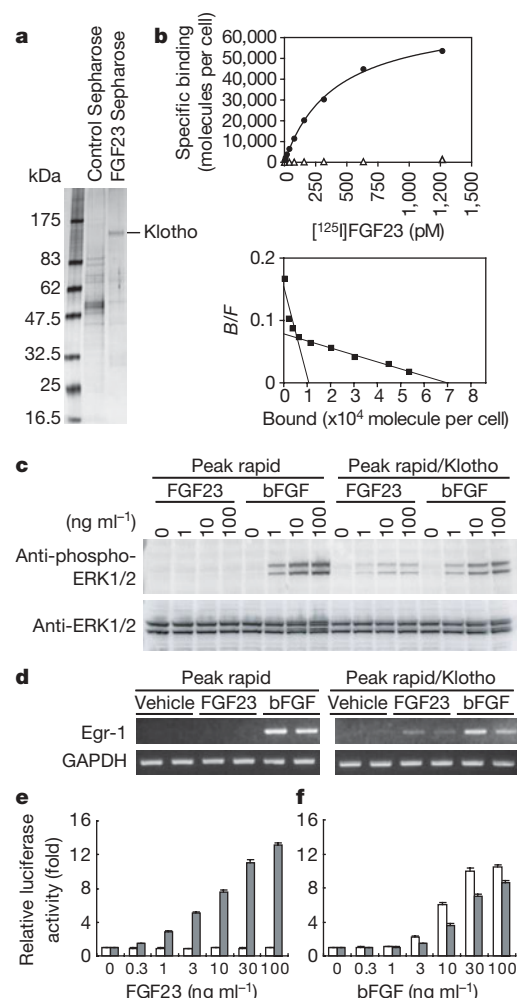


Figure 2 | FGF23 binds to Klotho and evokes cellular responses in Klotho-expressing cells. **a**, Klotho was determined as a major FGF23-binding protein from kidney. Among proteins recovered by FGF23-Sepharose from a murine kidney homogenate, Klotho was identified by MALDI-TOF mass spectrometry (see Methods). **b**, The specific binding of radiolabelled FGF23 to Klotho-expressing Chinese hamster ovary (CHO) cells was analysed by Scatchard plot analysis. Top: binding curve. Filled circles, CHO cells with Klotho; open triangles, CHO cells in the absence of Klotho. Bottom: Scatchard plot. For the steep (left-hand) portion, the parameters of the line are $K_d = 29 \text{ pM}$ and $B_{\text{max}} = 10,700$ molecules per cell; for the right-hand portion they are $K_d = 370 \text{ pM}$ and $B_{\text{max}} = 70,000$ molecules per cell. **c**, FGF23-induced, Klotho-dependent ERK activation was determined by immunoblotting using lysates of Klotho-expressing Peak rapid cells or control cells 10 min after treatment with FGF23 or bFGF. **d**, FGF23-induced *Egr-1* expression in Klotho-expressing Peak rapid cells or control cells was determined by treating the cells with vehicle or with 100 ng ml^{-1} FGF23 or bFGF for 30 min followed by RT-PCR ($n = 2$). GAPDH, glyceraldehyde-3-phosphate dehydrogenase. **e**, **f**, The signal transduction of FGF23 (**e**) or bFGF (**f**) in the presence (grey bars) or absence (white bars) of Klotho was assessed by the *Egr-1* reporter assay (see Methods) in Peak rapid cells. The results are expressed as means $\pm$ s.d. ($n = 3$).

after injection (Fig. 3c). Before these events, renal *CYP27B1* mRNA abundance increased and *CYP24* mRNA abundance decreased (Fig. 3f, g). A significant increase in the serum phosphate concentration was accompanied by an increased abundance of the type IIa sodium-phosphate co-transporter (NaPi2a) in the renal brush border membranes 24 h after injection (Fig. 3d, h). The serum calcium concentration increased and the serum PTH level tended to decrease in antibody-treated wild-type mice (Supplementary Table 2). All these changes caused by the antibody treatment were similar to alterations observed in Klotho-deficient (*kl/kl*) and FGF23-null mice. Serum FGF23 also increased, as in mice with impaired Klotho gene expressions (Fig. 3e). These findings indicate that the direct interaction between FGF23 and Klotho is essential for FGF23 function *in vivo*.

Although the importance of the interaction between Klotho and FGF23 for FGF23 signalling has been demonstrated both *in vitro* and

in vivo, an additional molecule in the FGF23 receptor system may be required because the tiny intracellular portion of Klotho seems too small to be able to evoke intracellular signalling. The widespread existence of this additional component was inferred from the fact that the Klotho-dependent FGF23 signalling was reproduced in cultured cells from a variety of origins (Supplementary Table 1). However, one exception—L6 rat myoblast cells, which lack endogenous FGF receptors²³ (FGFRs), failed to respond to FGF23 in the presence of Klotho—shed light on the possible involvement of FGFRs.

FGFRs are derived from four distinct genes denoted FGFR1–FGFR4 (refs 3, 24). The difference in the number of immunoglobulin-like loops (α -type and β -type) and alternative splicing of mRNA corresponding to the third immunoglobulin-like loop (IIIa, IIIb and IIIc) produce several subtypes of FGFRs. Interestingly, when various FGFR–Fc fusion proteins were added exogenously to cultures, only FGFR1 α (IIIc)–Fc or FGFR1 β (IIIc)–Fc significantly decreased the Klotho-dependent FGF23-induced activation of *Egr-1* promoter activity, whereas all types of FGFR–Fc protein could suppress the activation of *Egr-1* promoter activity by bFGF (Supplementary Fig. 5). To confirm the specific involvement of FGFR1(IIIc) in FGF23 signalling, we attempted to reconstitute the FGF23 receptor by expressing Klotho and various FGFRs in L6 cells. The co-expression of full-length FGFR1 β (IIIc) and Klotho in L6 cells enabled them to respond to FGF23 (Fig. 4a). Furthermore, phosphorylation of FGFR1 was observed in response to FGF23 in the presence of Klotho (Fig. 4b). FGFR substrate 2 (FRS2), which is a docking protein linking FGFRs to the Ras/MAPK pathway, and ERK1/2 were also phosphorylated by stimulation of FGF23 in the presence of Klotho.

To clarify the specific role of Klotho in the interaction of FGF23 and FGFR1(IIIc), we demonstrated specific binding of the extracellular domains of Klotho to FGFR1(IIIc) (Fig. 4c) and FGF23 (Fig. 4f). These interactions were independent of heparin. Heparin or other glycosaminoglycans were not able to substitute for Klotho in the specific binding of FGF23 to the cell surface of OK cells (Supplementary Fig. 2) and to FGFR1(IIIc) (Fig. 4d), and in the *Egr-1* promoter activation assay (Supplementary Fig. 6c), indicating the indispensability of Klotho for generating the specific FGF23–FGFR1(IIIc) interaction explained here. The addition of heparin or other glycosaminoglycans enhanced the activity of FGF23 significantly in the Klotho-dependent assay (Supplementary Fig. 6a, b) and stabilized the FGF23–Klotho–FGFR1(IIIc) complex effectively (Fig. 4d). Taken together, these results show that Klotho is fundamentally important for the high-affinity interaction between FGF23 and FGFR1(IIIc), and heparin or other glycosaminoglycans assist the stabilization of the FGF23–Klotho–FGFR1(IIIc) complex (Supplementary Discussion). In contrast to FGF23, bFGF competed with Klotho for FGFR1(IIIc) (Fig. 4e). The competition between Klotho and bFGF was also suggested by the data showing that Klotho attenuated bFGF activity in this study (Figs 2c, d, f and 4b). These findings indicate that FGFR1 is commonly used for bFGF and FGF23 signalling, and Klotho determines the ligand specificity in the FGFR1 usage. Klotho and FGFR1 are essential molecules that together comprise the FGF23 receptor, and this complex initiates cellular signalling by interacting with FGF23 at the cell surface.

We have revealed a physiological role for Klotho; namely, it converts FGFR1(IIIc), a canonical receptor for various FGFs, into a specific receptor for FGF23. This FGF23 receptor shows a strong affinity, one that is sufficient for interacting with physiological concentrations of FGF23. Moreover, the tissue-specific unique biological activity of FGF23 is likely to be regulated by the limited local distribution of Klotho. Given that Klotho, with its homologues β -Klotho²⁵ and KLPH (for Klotho–lactase-phlorizin hydrolase-related protein)²⁶, forms a family, other FGF–FGFR systems may be subject to similar conversions. This may be especially true for the other potential systemic FGFs that, together with FGF23, compose a unique FGF subfamily, namely FGF19 (ref. 27) and FGF21 (ref. 28). In fact, the

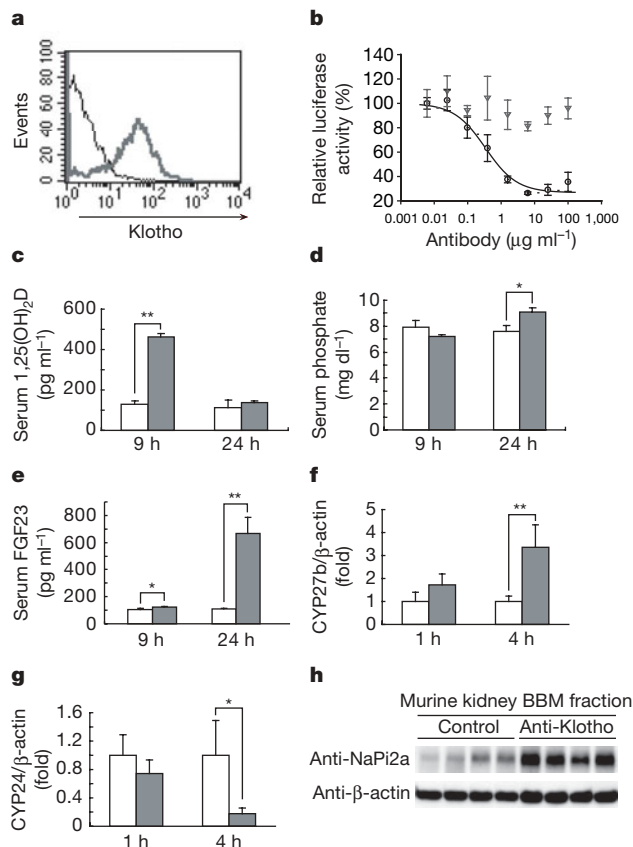


Figure 3 | An anti-Klotho monoclonal antibody antagonizes FGF23 both *in vitro* and *in vivo*. **a**, Specific binding of the anti-Klotho rat monoclonal antibody (70E2) to Klotho-expressing HEK-293 cells was measured by fluorescence-activated cell sorting analysis with the phycoerythrin-labelled anti-rat IgG antibody. Thin black line, HEK-293 cells; thick grey line, HEK-293 cells with Klotho. **b**, The 70E2 antibody inhibited FGF23 signalling *in vitro* in a dose-dependent manner, when 70E2 antibody (circles) or control IgG (triangles) was added with FGF23 (10 ng ml⁻¹) to Klotho-expressing cells in the *Egr-1* reporter assay. The results are expressed as means $\pm$ s.d. from an experiment performed in triplicate. **c–h**, The 70E2 antibody (3 mg kg⁻¹; grey bars) or vehicle (white bars) was injected into wild-type mice ($n = 5$ in each case). Serum concentrations of 1,25(OH)₂D (**c**), phosphate (**d**) and FGF23 (**e**) were determined 9 and 24 h after injection. The results are expressed as means $\pm$ s.e.m. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$, with Student's *t*-test. **f, g**, Renal expressions of CYP27B1 (1 α OHase) (**f**) and CYP24 (24OHase) (**g**) at 1 and 4 h after injection were determined by real-time RT-PCR ($n = 5$ in each case). The results are expressed as means $\pm$ s.d. Asterisk, $P < 0.05$; two asterisks, $P < 0.01$, with Student's *t*-test. **h**, Increased NaPi2a protein abundance in renal brush border membranes 24 h after injection was measured by western blotting ($n = 4$ in each case).

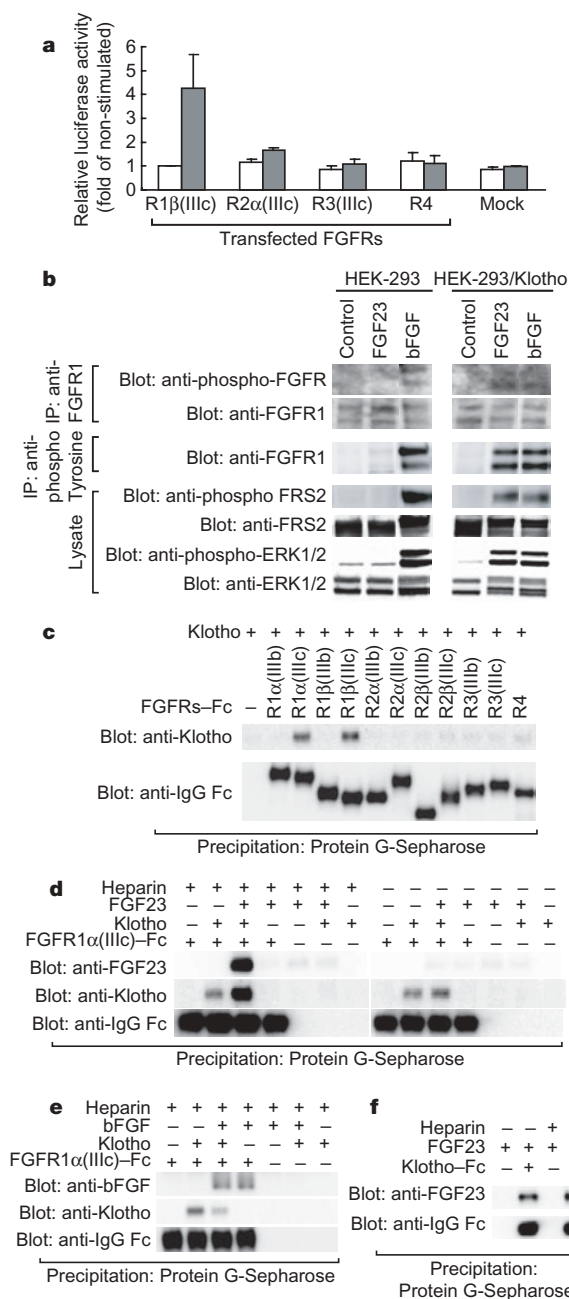


Figure 4 | Identification of FGFR1(IIIc) as a molecule that interacts with Klotho and is essential for FGF23 signalling. **a**, FGF23 signalling is restored in L6 cells by co-expression of Klotho and FGFR1 (IIIc). The Egr-1 reporter activation by FGF23 (100 ng ml⁻¹) was assessed in L6 cells that were transfected with various FGFRs with (grey bars) or without (white bars) the Klotho expression vector. The results are expressed as means $\pm$ s.d. ($n = 3$). **b**, FGF23 activated FGFR1, FRS2 and ERK1/2 in Klotho-expressing cells. Phosphorylation of FGFR1, FRS2 and ERK1/2 in the cell lysate prepared after treatment for 10 min with FGF23 or bFGF was determined by immunoprecipitation followed by immunoblotting with the indicated combination of antibodies (see Methods). **c**, Klotho and FGFR1(IIIc) interact specifically in the absence of heparin. The interaction between the extracellular domain of the Klotho protein and various FGFR-Fc fusion proteins were analysed by pull-down assay. **d–f**, Interactions between either FGF23 or bFGF, the extracellular domain of Klotho, Klotho-Fc, and FGFR1(IIIc)-Fc were analysed by pull-down assay. Heparin stabilized the FGF23–Klotho–FGFR1(IIIc) complex, whereas FGF23 was barely detectable in the absence of heparin (**d**). bFGF competed with Klotho for FGFR1(IIIc) (**e**). The interaction between FGF23 and Klotho was independent of heparin (**f**).

Klotho-dependent FGF23 signalling in this study did not stimulate thymidine incorporation (data not shown) despite the activation of FGFR1(IIIc) and the phosphorylation of ERK1/2. This non-mitogenic character is also documented with FGF21 (ref. 28). The impairment of bile acid metabolism in the recently developed β -Klotho KO mice²⁹ strongly implies the involvement of β -Klotho in the interaction between FGF19 and FGFR4 in regulating bile acid synthesis in the liver. Our discovery of this new type of receptor modulator may provide insights into the diversity of biological responses, especially those that are generated by receptor systems that have many similar, or apparently redundant, receptor and ligand subtypes.

METHODS

Detailed information on the methods used in this study is given in Supplementary Information.

FGF23-binding protein analysis. Purified mutant FGF23 (R176Q, R179Q) protein³⁰ was conjugated to *N*-hydroxysuccinimide-activated Sepharose. The 8,000g-unprecipitable and 100,000g-precipitable fraction of murine kidney homogenate was subjected to incubation with the FGF23-Sepharose resins. The proteins specifically retained on the resin were separated by SDS-PAGE and analysed by MALDI-TOF mass spectrometry.

Egr-1 reporter assay. The murine *Egr-1* promoter region (−1023 to +1)¹⁷ was cloned from murine genomic DNA and inserted into the pGL3 basic vector. Cells transfected with the *Egr-1* reporter were stimulated with recombinant FGF23 or bFGF on the following day. Heparin, other glycosaminoglycans, anti-Klotho antibodies or FGFR-Fc proteins were added together with the stimulants. The luciferase activity was measured after incubation for 24 h.

Immunoprecipitation and immunoblotting analysis. Immunoprecipitation was performed with anti-FGFR1 antibody-conjugated or anti-phosphotyrosine antibody-conjugated beads. Immunoblotting was performed with anti-Egr-1, anti-phospho-ERK, anti-ERK, anti-FGFR1, anti-phospho-FGFR1, anti-FRS2 or anti-phospho-FRS2 antibody.

Pull-down assay. Various proteins were incubated with 0.5 µg of an FGFR1(IIIc)-Fc fusion protein or conditioned medium containing the recombinant murine Klotho-Fc fusion with or without heparin. Proteins adsorbed by Protein G-Sepharose were removed and analysed by western blotting for FGFR-Fc, Klotho, FGF23 or bFGF.

Received 12 April; accepted 9 October 2006.

Published online 29 October 2006.

1. The ADHR Consortium. Autosomal dominant hypophosphataemic rickets is associated with mutations in FGF23. *Nature Genet.* **26**, 345–348 (2000).
2. Shimada, T. *et al.* Cloning and characterization of FGF23 as a causative factor of tumor-induced osteomalacia. *Proc. Natl Acad. Sci. USA* **98**, 6500–6505 (2001).
3. Itoh, N. & Ornitz, D. M. Evolution of the *Fgf* and *Fgfr* gene families. *Trends Genet.* **20**, 563–569 (2004).
4. Yamazaki, Y. *et al.* Increased circulatory level of biologically active full-length FGF-23 in patients with hypophosphatemic rickets/osteomalacia. *J. Clin. Endocrinol. Metab.* **87**, 4957–4960 (2002).
5. Jonsson, K. B. *et al.* Fibroblast growth factor 23 in oncogenic osteomalacia and X-linked hypophosphatemia. *N. Engl. J. Med.* **348**, 1656–1663 (2003).
6. Shimada, T. *et al.* FGF-23 is a potent regulator of vitamin D metabolism and phosphate homeostasis. *J. Bone Miner. Res.* **19**, 429–435 (2004).
7. Kuro-o, M. *et al.* Mutation of the mouse *klotho* gene leads to a syndrome resembling ageing. *Nature* **390**, 45–51 (1997).
8. Riminucci, M. *et al.* FGF-23 in fibrous dysplasia of bone and its relationship to renal phosphate wasting. *J. Clin. Invest.* **112**, 683–692 (2003).
9. Murer, H., Hernando, N., Forster, I. & Biber, J. Proximal tubular phosphate reabsorption: molecular mechanisms. *Physiol. Rev.* **80**, 1373–1409 (2000).
10. Wikvall, K. Cytochrome P450 enzymes in the bioactivation of vitamin D to its hormonal form. *Int. J. Mol. Med.* **7**, 201–209 (2001).
11. Gashler, A. & Sukhatme, V. P. Early growth response protein 1 (Egr-1): prototype of a zinc-finger family of transcription factors. *Prog. Nucleic Acid Res. Mol. Biol.* **50**, 191–224 (1995).
12. Liu, L., Tsai, J. C. & Aird, W. C. Egr-1 gene is induced by the systemic administration of the vascular endothelial growth factor and the epidermal growth factor. *Blood* **96**, 1772–1781 (2000).
13. Nabeshima, Y. Klotho: a fundamental regulator of aging. *Ageing Res. Rev.* **1**, 627–638 (2002).
14. Imura, A. *et al.* Secreted Klotho protein in sera and CSF: implication for post-translational cleavage in release of Klotho protein from cell membrane. *FEBS Lett.* **565**, 143–147 (2004).
15. Yu, X. *et al.* Analysis of the biochemical mechanisms for the endocrine actions of fibroblast growth factor-23. *Endocrinology* **146**, 4647–4656 (2005).

16. Yamashita, T., Konishi, M., Miyake, A., Inui, K. & Itoh, N. Fibroblast growth factor (FGF)-23 inhibits renal phosphate reabsorption by activation of the mitogen-activated protein kinase pathway. *J. Biol. Chem.* **277**, 28265–28270 (2002).
 17. Christy, B. A., Lau, L. F. & Nathans, D. A gene activated in mouse 3T3 cells by serum growth factors encodes a protein with 'zinc finger' sequences. *Proc. Natl Acad. Sci. USA* **85**, 7857–7861 (1988).
 18. Shimada, T. *et al.* Targeted ablation of Fgf23 demonstrates an essential physiological role of FGF23 in phosphate and vitamin D metabolism. *J. Clin. Invest.* **113**, 561–568 (2004).
 19. Sitara, D. *et al.* Homozygous ablation of fibroblast growth factor-23 results in hyperphosphatemia and impaired skeletogenesis, and reverses hypophosphatemia in PheX-deficient mice. *Matrix Biol.* **23**, 421–432 (2004).
 20. Yoshida, T., Fujimori, T. & Nabeshima, Y. Mediation of unusually high concentrations of 1,25-dihydroxyvitamin D in homozygous klotho mutant mice by increased expression of renal 1 α -hydroxylase gene. *Endocrinology* **143**, 683–689 (2002).
 21. Tsujikawa, H., Kurotaki, Y., Fujimori, T., Fukuda, K. & Nabeshima, Y. Klotho, a gene related to a syndrome resembling human premature aging, functions in a negative regulatory circuit of vitamin D endocrine system. *Mol. Endocrinol.* **17**, 2393–2403 (2003).
 22. Larsson, T., Nisbeth, U., Ljunggren, O., Juppner, H. & Jonsson, K. B. Circulating concentration of FGF-23 increases as renal function declines in patients with chronic kidney disease, but does not change in response to variation in phosphate intake in healthy volunteers. *Kidney Int.* **64**, 2272–2279 (2003).
 23. Vainikka, S. *et al.* Signal transduction by fibroblast growth factor receptor-4 (FGFR-4). Comparison with FGFR-1. *J. Biol. Chem.* **269**, 18320–18326 (1994).
 24. Ornitz, D. M. & Itoh, N. Fibroblast growth factors. *Genome Biol.* **2**, 3005.1–3005.12 (2001).
 25. Ito, S. *et al.* Molecular cloning and expression analyses of mouse *betaklotho*, which encodes a novel Klotho family protein. *Mech. Dev.* **98**, 115–119 (2000).
 26. Ito, S., Fujimori, T., Hayashizaki, Y. & Nabeshima, Y. Identification of a novel mouse membrane-bound family 1 glycosidase-like protein, which carries an atypical active site structure. *Biochim. Biophys. Acta* **1576**, 341–345 (2002).
 27. Holt, J. A. *et al.* Definition of a novel growth factor-dependent signal cascade for the suppression of bile acid biosynthesis. *Genes Dev.* **17**, 1581–1591 (2003).
 28. Kharitonov, A. *et al.* FGF-21 as a novel metabolic regulator. *J. Clin. Invest.* **115**, 1627–1635 (2005).
 29. Ito, S. *et al.* Impaired negative feedback suppression of bile acid synthesis in mice lacking betaKlotho. *J. Clin. Invest.* **115**, 2202–2208 (2005).
 30. Shimada, T. *et al.* Mutant FGF-23 responsible for autosomal dominant hypophosphatemic rickets is resistant to proteolytic cleavage and causes hypophosphatemia *in vivo*. *Endocrinology* **143**, 3179–3182 (2002).
- Supplementary Information** is linked to the online version of the paper at www.nature.com/nature. A figure summarizing the main result of this paper is available in Supplementary Information.
- Acknowledgements** We thank R. Imai, J. Murakami, M. Sato-Fujisawa, R. Hino, K. Sakuma, K. Ono and N. Yoshii for technical assistance, and Y. Nabeshima, T. Yoneya, T. Muto, T. Kawata, Y. Aono, J. Yasutake, N. Kasai and other members of our laboratories for advice and discussion. This work was supported in part by grants from the Ministry of Education, Culture, Sports, Science and Technology, and from the Ministry of Health, Labour, and Welfare, Japan.
- Author Information** Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to I.U. (iurakawa@kirin.co.jp).

Phospholipids and the origin of cationic gating charges in voltage sensors

Daniel Schmidt^{1*}, Qiu-Xing Jiang^{1*} & Roderick MacKinnon¹

Cells communicate with their external environment through physical and chemical processes that take place in the cell-surrounding membrane. The membrane serves as a barrier as well as a special environment in which membrane proteins are able to carry out important processes. Certain membrane proteins have the ability to detect the membrane voltage and regulate ion conduction or enzyme activity^{1,2}. Such voltage-dependent processes rely on the action of protein domains known as voltage sensors, which are embedded inside the cell membrane and contain an excess of positively charged amino acids, which react to an electric field. How does the membrane create an environment suitable for voltage sensors? Here we show under a variety of conditions that the function of a voltage-dependent K⁺ channel is dependent on the negatively charged phosphodiester of phospholipid molecules. A non-voltage-dependent K⁺ channel does not exhibit the same dependence. The data lead us to propose that the phospholipid membrane, by providing stabilizing interactions between positively charged voltage-sensor arginine residues and negatively charged lipid phosphodiester groups, provides an appropriate environment for the energetic stability and operation of the voltage-sensing machinery. We suggest that the usage of arginine residues in voltage sensors is an adaptation to the phospholipid composition of cell membranes.

Structural studies on the voltage-dependent K⁺ channels KvAP and Kv1.2 show that the voltage sensors are located at the protein–lipid interface and that arginine residues may be exposed to the membrane^{3–8}. In the structure of the Kv1.2 K⁺ channel, for example, the sensors are arranged as nearly independent domains, with much of their surface surrounded by lipid. The atomic structures have led to the hypothesis that the interaction of voltage sensors with the membrane is important not only to their structure but also to their function^{3–8}. This hypothesis is testable through studies on the effect of membrane lipid composition on voltage-dependent channel function.

Figure 1 shows KvAP K⁺ channels in three different experimental conditions. In each condition the membrane voltage was held at –100 mV and then stepped to more positive depolarizing voltages to

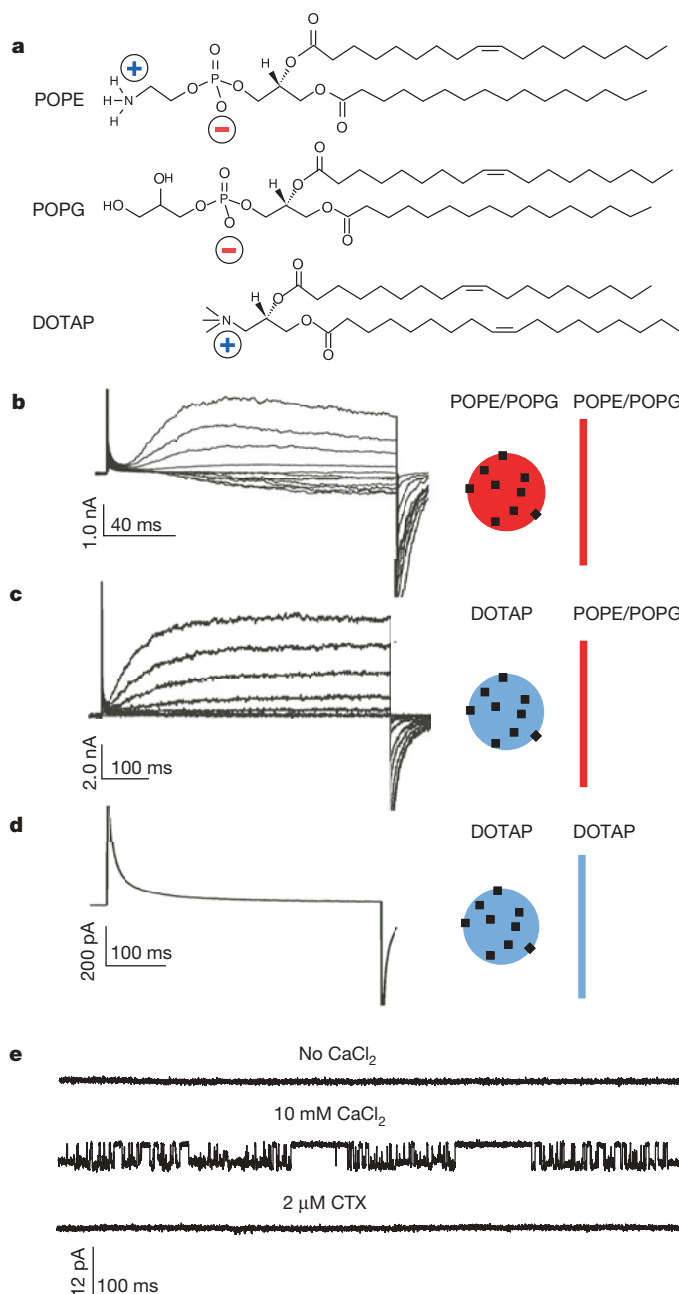


Figure 1 | Assessment of KvAP and MthK function in phospholipid and DOTAP membranes. **a**, Chemical structures of POPE, POPG and DOTAP. **b**, KvAP (black squares in cartoon) in 3:1 POPE:POPG vesicles (red circle) fused into a 3:1 POPE:POPG bilayer (red line) yielded voltage-dependent currents (left traces). **c**, **d**, KvAP in DOTAP vesicles (blue circle) gave functional channels (traces in **c**) after fusion into a POPE:POPG bilayer (red line), but failed to function (trace in **d**) after fusion into a DOTAP bilayer (blue line). Voltage pulses in **b** and **c**: –100 to 40 mV, $\Delta V = 10$ mV, holding potential (h.p.) is –100 mV. The pulse in **d** was from h.p. –100 to 100 mV. **e**, Single-channel recordings of MthK in DOTAP membranes with no CaCl₂ (top), 10 mM CaCl₂ (middle), and after the addition of 2.0 μM CTX (bottom). h.p. is –150 mV, and channels open downward.

¹Howard Hughes Medical Institute, Laboratory of Molecular Neurobiology and Biophysics, Rockefeller University, 1230 York Avenue, New York, New York 10021, USA.

*These authors contributed equally to this work.

open the channels. When KvAP channels were present in phospholipid membranes made of 1-palmitoyl-2-oleoyl-glycero-3-phosphoethanolamine (POPE) and 1-palmitoyl-2-oleoyl-glycero-3-phospho-1-glycerol (POPG) the channels opened in a voltage-dependent manner (Fig. 1b, c). By contrast, when channels were present in membranes consisting of a lipid known as 1,2-dioleoyl-3-trimethylammonium propane (DOTAP)⁹, which contains a positively charged trimethylammonium group instead of a negatively charged phosphodiester (Fig. 1a), no channel activity was observed (Fig. 1d). The recording in Fig. 1c shows that KvAP channels are able to reside in a lipid membrane consisting of only DOTAP because the channels were first reconstituted into pure DOTAP lipid vesicles, which were then fused with the planar membrane of phospholipids. The presence of functional channels in the planar phospholipid membrane in Fig. 1c means that KvAP channels had to be present in the DOTAP lipid vesicles that were used to deliver channels to the planar membrane.

Figure 1e shows that a Ca^{2+} -dependent K^+ channel, MthK, is able to function in pure DOTAP membranes. MthK has a pore that is similar to that of KvAP but contains a gating ring structure in the cytoplasm (instead of voltage sensors in the membrane) that enables intracellular Ca^{2+} to open the pore¹⁰. In DOTAP membranes MthK is activated by Ca^{2+} and inhibited by the scorpion toxin charybdotoxin (CTX) (Fig. 1e). This control experiment shows that DOTAP vesicles can indeed fuse with DOTAP planar membranes and that DOTAP membranes do not prevent the pore of a K^+ channel from function-

ing. The control also implies that the inability of DOTAP membranes to support KvAP function is related to the voltage-sensing mechanism.

If KvAP channels are present but silent in DOTAP planar membranes then it might be possible to activate them through the subsequent addition of phospholipids to the membranes. Several minutes after fusing channel-containing DOTAP vesicles with DOTAP planar membranes, empty phospholipid vesicles (without KvAP) were fused, resulting in channel activity (Fig. 2a). Because the phospholipid vesicles themselves did not contain KvAP channels, the channels were presumably present in the DOTAP planar membrane but were not functional. In these experiments the number of active channels was always small and in several respects their function was not identical to channels in pure phospholipid membranes. The important point, however, is that channels began to function in a voltage-dependent manner only after phospholipids were delivered to the membrane: that is, in the context of a DOTAP planar membrane KvAP is a 'phospholipid-activated' voltage-dependent K^+ channel.

If DOTAP membranes fail to support channel function then we might expect to observe channels exhibiting 'abnormal' function (that is intermediate behaviour between normal and nonfunctional) in composite membranes containing DOTAP and phospholipids. We therefore studied the effect on voltage-dependent gating of systematically decreasing the molar ratio of phospholipids (Fig. 2b). A dilute solution of phospholipid vesicles containing KvAP channels at

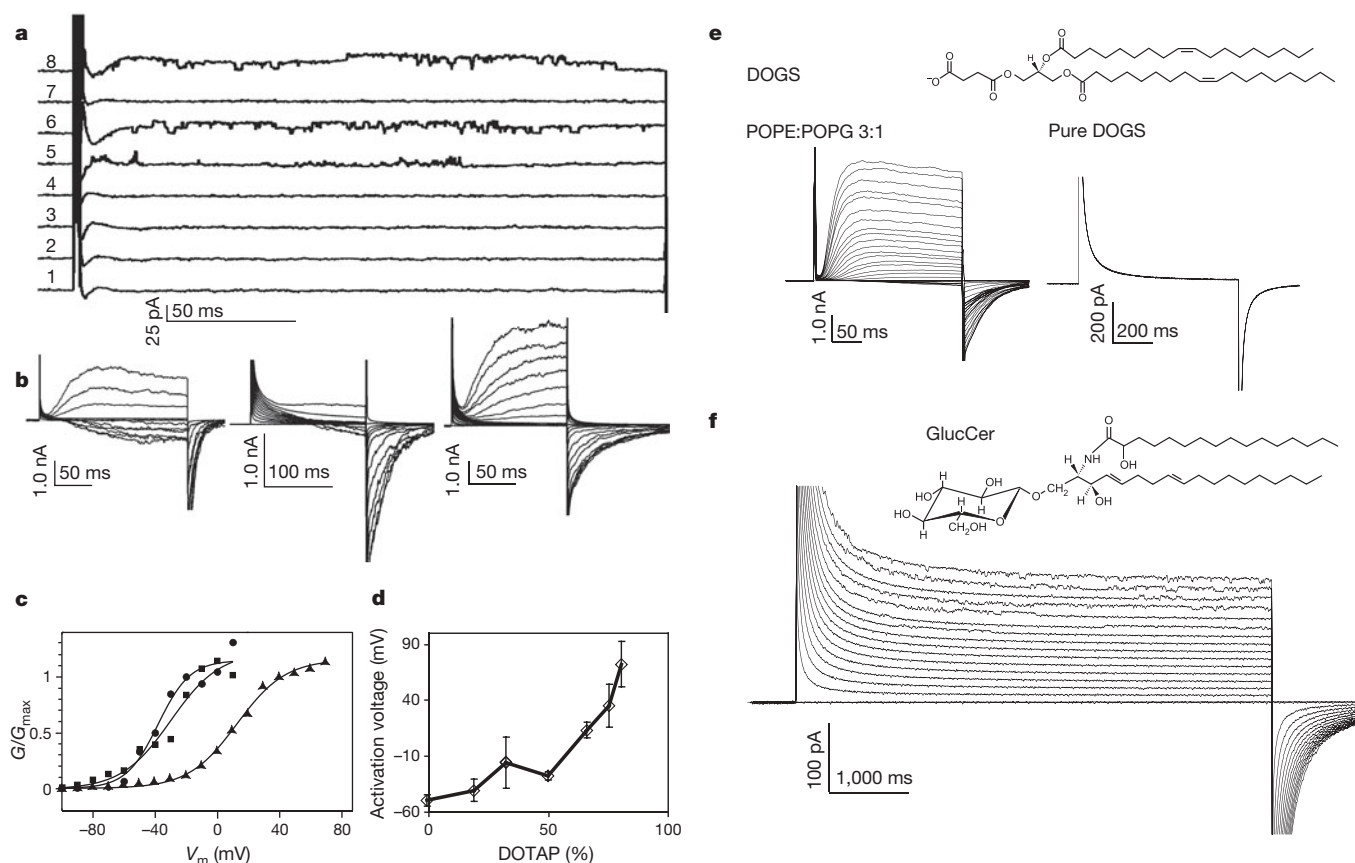


Figure 2 | KvAP function depends on membrane lipid composition. **a**, KvAP-containing DOTAP vesicles were fused into a DOTAP bilayer. Channel activity was monitored with pulses from -100 to 100 mV every two minutes. Empty POPE:POPG vesicles were introduced between traces 3 and 4. A triple-exponential function was used to subtract capacitance transients. Traces 2–8 were shifted upward successively by 50 pA. **b**, KvAP in POPE:POPG vesicles at a protein-to-lipid ratio (w/w) of 1.0 was fused into bilayers of mixed DOTAP and POPE:POPG with the percentage of DOTAP 0% (left), 50% (middle) and 67% (right). Voltage pulses: h.p. -100 mV to 10 mV (left and middle) and h.p. -100 mV to 70 mV (right), $\Delta V = 10$ mV. **c**, Boltzmann functions (solid lines) fit data from **b** (normalized conductance $G/G_{\max}$ as a function of membrane voltage V_m) with the midpoint activation voltage $V_{0.5}$ (mV) and electrical valence Z : -42 , 3.1 (0% , circle); -25 , 1.9 (50% , square); 14 , 1.7 (67% , triangle). **d**, $V_{0.5}$ (mean $\pm$ s.e.m. or range of mean, $n = 2-5$) versus percentage of DOTAP. **e**, KvAP in DOGS vesicles fused into bilayers of POPE:POPG (left) and DOGS (right). Voltage pulses: -80 to $+180$ mV, $\Delta V = 10$ mV, h.p. -100 mV (left); from h.p. -100 to 120 mV (right). **f**, KvAP in DOTAP vesicles fused into a GlucCer bilayer. Voltage pulses: -150 to 150 mV, $\Delta V = 20$ mV, h.p. -150 mV. Structure of GlucCer given above trace.

a high protein-to-lipid ratio was used for fusion: this condition best ensured that after fusion the lipid composition of the planar membrane was dominated by the planar membrane lipids (the phospholipid-DOTAP mixture) rather than the vesicle lipids. As the molar ratio of phospholipids was decreased, successively larger depolarization voltages were required to open the channels, and the activation curve became less steep (Fig. 2b, c). At present, we do not know what channel conformations are favoured by the depletion of phospholipids and replacement with DOTAP (that is closed, inactivated, or a non-native state). It is clear, however, that gating becomes progressively more abnormal as the mole fraction of phospholipids is decreased (Fig. 2b, c). The absence of function in pure DOTAP membranes (Fig. 1d) appears to represent the limit of the dilution experiment (Fig. 2b–d), and the restoration of voltage-dependent activity upon addition of phospholipids (Fig. 2a) appears to reflect the fulfilment of a phospholipid requirement for voltage-dependent channel function.

The inability of DOTAP to support KvAP function could be related to the absence of a negatively charged phosphodiester or to the presence of a positively charged trimethylammonium group (Fig. 1a). To distinguish these two possibilities we studied additional non-phospholipid membranes, 1,2-dioleoyl-glycero-3-succinate (DOGS) and soy glucocerebrosides (GlucCer), which do not have a trimethylammonium group (Fig. 2e, f). DOGS is a carboxyl-containing anionic lipid at neutral pH. It forms large unilamellar vesicles in which reconstituted KvAP channels are visible under cryo-electron microscopy (Supplementary Fig. 1) and functional when fused into POPE:POPG planar membranes (Fig. 2e). DOGS also forms stable planar membranes; however, no channel activity is observed. GlucCer is a neutral lipid with a ceramide backbone and sugar head group. Rare (one to three channels) opening events are observed at very positive membrane voltages (Fig. 2f). This lipid is heterogeneous, contains trace phosphate in our phosphate assays, and may exhibit complex phase behaviour. We cannot be certain whether GlucCer or contaminant lipids support the rare openings at very positive voltages. In either case, KvAP function in this membrane is very abnormal. DOTAP, DOGS and GlucCer are chemically very different from each other. None of them appear to create a suitable environment for normal KvAP function.

What special property of phospholipid membranes favours voltage-dependent channel function? We examined the head group requirement by supplementing DOTAP membranes with the zwitterionic phospholipids 1,2-dioleoyl-glycero-3-phosphocholine (DOPC), 1,2-dimyristoyl-glycero-3-phosphocholine (DMPC) and 1,2-didodecyl-glycero-3-phosphocholine (DDECPC), which have successively shorter acyl chains (Fig. 3a). DMPC and DDECPC alone do not form stable planar membranes in the decane bilayer system. Voltage-dependent channel function was supported using phospholipids with acyl chains as short as ten carbons (DDECPC). These short chain lipids apparently plug into the DOTAP membrane and fulfill a head group requirement of the channel. A lipid that supplies only the phosphate group, 1,2-dioleoyl-glycerol-3-phosphate (DOPA), whether mixed with DOTAP or alone, also supports KvAP function (Fig. 3b, Supplementary Fig. 2). Thus, the phosphate group appears to be important for voltage-dependent channel function.

The behaviour of KvAP in mixed phospholipid membranes consisting of 1-palmitoyl-2-oleoyl-glycero-3-phosphocholine (POPC) and POPE or POPG is shown (Fig. 3c, d). POPC:POPG membranes (Fig. 3c) are similar to POPE:POPG membranes (Fig. 1b and Fig. 2b) in that both contain a zwitterionic (often referred to as neutral) phospholipid (POPE or POPC) and an anionic phospholipid (POPG). However, these membranes differ in that POPE and POPC have different tendencies to curve¹¹. KvAP functions similarly in both membranes, with midpoints of activation of approximately -40 mV for POPE:POPG membranes and -30 mV for POPC:POPG membranes. The POPE:POPC membrane contains only zwitterionic phospholipids: here the midpoint of KvAP activation is shifted to

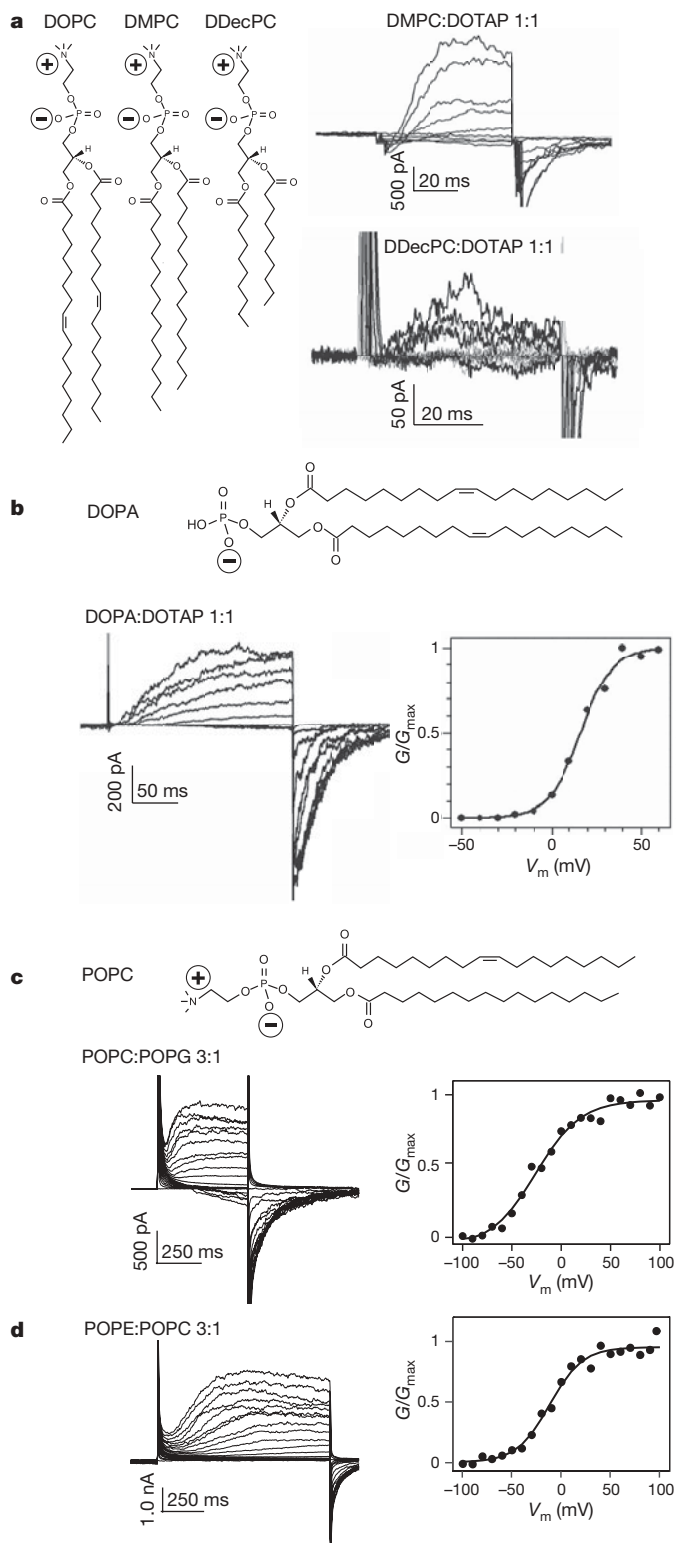


Figure 3 | Phosphate group is important for KvAP voltage-dependent gating. **a**, KvAP in DOTAP vesicles fused into bilayers of 1:1 DOTAP/DMPC (a 14-carbon acyl chain) or DOTAP/DDECPC (a ten-carbon acyl chain). Voltage pulses: -100 to 100 mV, $\Delta V = 20$ mV, h.p. = -100 mV. **b**, KvAP in DOPA vesicles (protein-to-lipid ratio 1.0 w/w) fused into bilayers of 1:1 DOTAP/DOPA. Voltage pulses: h.p. -80 to 60 mV, $\Delta V = 10$ mV. Boltzmann function (continuous line in the right panel) fitted to normalized tail currents (black dots) gave $V_{0.5}$ (mV) and Z : 18, 2.5. **c**, **d**, KvAP in POPC vesicles fused into membranes of 3:1 POPC/POPG (**c**), and POPE/POPC (**d**). Voltage pulses: -100 to 100 mV, $\Delta V = 10$ mV, h.p. -100 mV. Boltzmann functions (continuous lines in right panels) yielded $V_{0.5}$ (mV) and Z : -30.5, 1.8 (**c**), and -9.5, 1.6 (**d**). CTX was used for capacitance transient subtraction in **a** and **b**.

approximately -10 mV, but voltage-dependent function is clearly intact (Fig. 3d). Other voltage-dependent channels have also been shown to function in purely zwitterionic phospholipid membranes^{12–14}. The comparison of KvAP function in POPE:POPG, POPC:POPG, and POPE:POPC membranes leads us to conclude that membrane curvature tendency and net charge may have modest influences on KvAP function. However, the presence or absence of the phosphate group appears to have a more dramatic influence on voltage-dependent channel function.

In Fig. 4 we begin to investigate the chemical properties of the phosphate group that are important. The phospholipid 1,2-dioleoyl-glycero-3-ethylphosphocholine (EDOPC) is similar to DOPC with respect to many of its physical properties¹⁵. However, the ethylation means that the resulting phosphotriester is uncharged and it

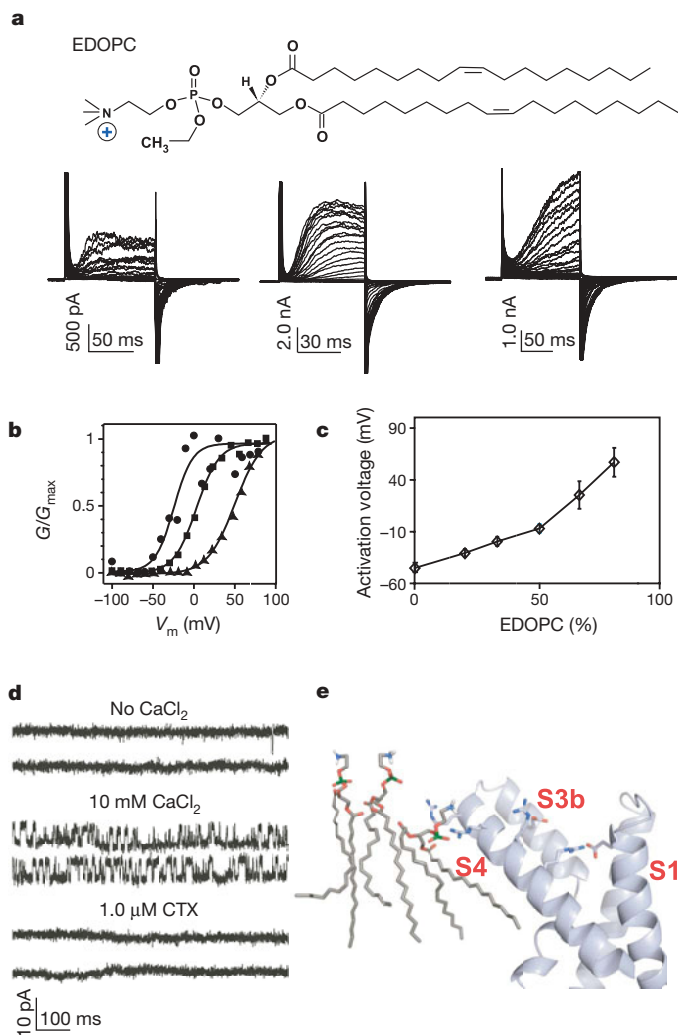


Figure 4 | Role of the negative charge on the phosphate group. **a**, KvAP in POPE:POPG vesicles (protein-to-lipid ratio 1.0 w/w) fused into bilayers of mixed EDOPC and POPE:POPG with percentage of EDOPC 33% (left), 66% (middle) and 80% (right). Structure of EDOPC shown above. Voltage pulses: -100 to 100 mV, $\Delta V = 5$ or 10 mV, h.p. -100 mV. **b**, Boltzmann functions (continuous lines) from normalized tail currents (black dots) measured from traces in **a** with $V_{0.5}$ (mV) and Z : -25 , 2.0 (33%, circle); 19 , 1.7 (66%, square); and 70 , 1.5 (80%, triangle). **c**, Activation voltage $V_{0.5}$ (mean $\pm$ s.e.m., $n = 3$) versus percentage of EDOPC. **d**, Single-channel activity of MthK in EDOPC membranes with no CaCl_2 (top), 10 mM CaCl_2 (middle), and after addition of 1.0 μM CTX (bottom). h.p. -100 mV. Channels open downward. **e**, The hypothesized interaction between the side chains of Arg residues (first two) in the voltage sensor (Protein Data Bank code 1ORS) and lipid phosphodiester groups (green and red). Voltage-sensor α -helices S1, S3b and S4 are labelled for orientation purposes.

does not participate in intermolecular hydrogen bonding with itself (Fig. 4a)¹⁵. When normal phospholipids (POPE:POPG) are diluted by increasing the mole fraction of EDOPC, the effects on KvAP gating are qualitatively similar to those observed in DOTAP membranes (Fig. 2b–d and Fig. 4a–c). No KvAP activity is observed in membranes consisting of 95% EDOPC and 5% DOTAP. In control experiments with EDOPC membranes the MthK channel conducts K^+ , is Ca^{2+} -activated, and inhibited by CTX (Fig. 4d). Therefore EDOPC membranes support the function of a non-voltage-dependent K^+ channel. It appears that the negative charge on the lipid phosphate group and possibly its hydrogen-bonding potential is specifically important for the function of the KvAP K^+ channel.

KvAP channels function in a variety of phospholipid membranes, including many that have not been presented here. Rates of opening and closing and midpoints of activation vary among membranes of different composition, but voltage-dependent gating is fundamentally intact. Among the many important properties of a lipid bilayer, including its tendency to curve and whether it is net charged or zwitterionic, the presence of the anionic phosphate group appears to be particularly important for KvAP function. This observation suggests that the lipid membrane might provide an environment that is suitable for voltage sensors because phosphate groups can serve as counter-charges for arginine residues, as depicted (Fig. 4e). Phosphate groups could be made available through specific binding of lipids to a channel, as appears to be the case for sphingomyelin and certain voltage-dependent channels¹⁶, or through unbound phospholipids in the vicinity of the channel. Arginine–phosphate-group interactions have been described in recent molecular dynamics simulations, which show local membrane distortions that enable the phosphate groups (and water molecules) to approach arginine residues below the surface¹⁷.

Further experiments are needed to confirm the hypothesis that direct interactions between arginine side chains and lipid phosphodiester stabilize the voltage sensor. The positive charge and multidentate hydrogen-bonding capacity of arginine's guanidinium group makes it an excellent chemical match for favourable electrostatic and hydrogen bonding interactions with the lipid phosphodiester. We suggest that the usage of positively charged amino acids with a preference towards arginine in voltage sensors is an adaptation to the phospholipid composition of the cell membrane.

METHODS

Please refer to the Supplementary Information for details.

Preparation and reconstitution of KvAP and MthK channels. KvAP was expressed and purified following published procedures¹⁸, and was concentrated to 5 – 10 mg ml^{-1} before reconstitution. The reconstitution followed a modification of a published procedure¹⁹. Lipids of desired compositions were prepared and solubilized (partially or completely) with detergents. The protein and the lipids were mixed at ratios ranging from 0.1 to 1.0 . The detergents were then removed by dialysis. Preparation and reconstitution of MthK was based on ref. 10.

Electrophysiological recordings. Bilayer experiments followed published procedures¹⁸. Bilayers were formed over a 300 μm hole in a polystyrene partition that separated two aqueous chambers²⁰. After vesicles with channels were fused into the bilayer, voltage-clamp measurements in whole-cell mode were made. All voltages are reported according to electrophysiological convention, with the extracellular side of the channel taken as ground.

For single-channel recordings of MthK in DOTAP or EDOPC membranes, 5 – 10 mM CaCl_2 was added to the side corresponding to the intracellular side of the channel. Charybdotoxin was added to the side corresponding to the extracellular side of the channel.

Received 5 September; accepted 3 November 2006.

Published online 29 November 2006.

1. Sigworth, F. J. Voltage gating of ion channels. *Q. Rev. Biophys.* **27**, 1–40 (1994).
2. Murata, Y., Iwasaki, H., Sasaki, M., Inaba, K. & Okamura, Y. Phosphoinositide phosphatase activity coupled to an intrinsic voltage sensor. *Nature* **435**, 1239–1243 (2005).
3. Jiang, Y. *et al.* X-ray structure of a voltage-dependent K^+ channel. *Nature* **423**, 33–41 (2003).
4. Jiang, Y., Ruta, V., Chen, J., Lee, A. & MacKinnon, R. The principle of gating charge movement in a voltage-dependent K^+ channel. *Nature* **423**, 42–48 (2003).

5. Lee, S. Y., Lee, A., Chen, J. & MacKinnon, R. Structure of the KvAP voltage-dependent K⁺ channel and its dependence on the lipid membrane. *Proc. Natl Acad. Sci. USA* **102**, 15441–15446 (2005).
6. Long, S. B., Campbell, E. B. & MacKinnon, R. Crystal structure of a mammalian voltage-dependent Shaker family K⁺ channel. *Science* **309**, 897–903 (2005).
7. Long, S. B., Campbell, E. B. & MacKinnon, R. Voltage sensor of Kv1.2: structural basis of electromechanical coupling. *Science* **309**, 903–908 (2005).
8. Jiang, Q. X., Wang, D. N. & MacKinnon, R. Electron microscopic analysis of KvAP voltage-dependent K⁺ channels in an open conformation. *Nature* **430**, 806–810 (2004).
9. Campbell, R. B., Balasubramanian, S. V. & Straubinger, R. M. Phospholipid–cationic lipid interactions: influences on membrane and vesicle properties. *Biochim. Biophys. Acta* **1512**, 27–39 (2001).
10. Jiang, Y. *et al.* Crystal structure and mechanism of a calcium-gated potassium channel. *Nature* **417**, 515–522 (2002).
11. Israelachvili, J. N. *Intermolecular and Surface Forces* 2nd edn, 366–394 (Academic Press, London, 1991).
12. Hui, K., McIntyre, D. & French, R. J. Conotoxins as sensors of local pH and electrostatic potential in the outer vestibule of the sodium channel. *J. Gen. Physiol.* **122**, 63–79 (2003).
13. MacKinnon, R., Latorre, R. & Miller, C. Role of surface electrostatics in the operation of a high-conductance Ca²⁺-activated K⁺ channel. *Biochemistry* **28**, 8092–8099 (1989).
14. Sun, T., Naini, A. A. & Miller, C. High-level expression and functional reconstitution of Shaker K⁺ channels. *Biochemistry* **33**, 9992–9999 (1994).
15. MacDonald, R. C. *et al.* Physical and biological properties of cationic triesters of phosphatidylcholine. *Biophys. J.* **77**, 2612–2629 (1999).
16. Ramu, Y., Xu, Y. & Lu, Z. Enzymatic activation of voltage-gated potassium channels. *Nature* **442**, 696–699 (2006).
17. Freites, J. A., Tobias, D. J., von Heijne, G. & White, S. H. Interface connections of a transmembrane voltage sensor. *Proc. Natl Acad. Sci. USA* **102**, 15059–15064 (2005).
18. Ruta, V., Jiang, Y., Lee, A., Chen, J. & MacKinnon, R. Functional analysis of an archaeobacterial voltage-dependent K⁺ channel. *Nature* **422**, 180–185 (2003).
19. Heginbotham, L., LeMasurier, M., Kolmakova-Partensky, L. & Miller, C. Single *Streptomyces lividans* K⁺ channels: functional asymmetries and sidedness of proton activation. *J. Gen. Physiol.* **114**, 551–560 (1999).
20. Miller, C. *Ion Channel Reconstitution* (Plenum Press, New York, 1986).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank Y. Jiang for the MthK channel construct and A. Lee for the CTX. This work was supported by an NIH grant to R.M. R.M. is an investigator in the Howard Hughes Medical Institute.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.M. (mackinn@rockefeller.edu).

CORRIGENDUM

doi:10.1038/nature05421

Smad4 signalling in T cells is required for suppression of gastrointestinal cancer

Byung-Gyu Kim, Cuiling Li, Wenhui Qiao, Mizuko Mamura, Barbara Kasprzak, Miriam Anver, Lawrence Wolfrim, Suntaek Hong, Elizabeth Mushinski, Michael Potter, Seong-Jin Kim, Xin-Yuan Fu, Chuxia Deng & John J. Letterio

Nature 441, 1015–1019 (2006)

In this Letter, the surname of co-author Barbara Kasprzak was misspelled as Kaspercak.

CORRIGENDUM

doi:10.1038/nature05424

Definitive fossil evidence for the extant avian radiation in the CretaceousJulia A. Clarke, Claudia P. Tambussi, Jorge I. Noriega,
Gregory M. Erickson & Richard A. KetchamNature 433, 305–308 (2005)

We wish to correct an oversight noted¹ in one line of our systematic paleontology section, altering this line from “*Vegavis iaa*i, new taxon” to “*Vegavis iaa*i, new taxa” to indicate our intent to name two taxa: one that under the current International Code of Zoological Nomenclature (ICZN)² is to be at the rank of a genus (that is, *Vegavis*), and the second that is to be at the rank of species (that is, *iaai*). Article 13.4 of this Code² specifies that the “combined description or definition of a new nominal genus or subgenus and a single included new nominal species, if marked by “gen. nov., sp. nov.” or an equivalent expression, is deemed to confer availability on each name under Article 13.1.1”. This line might also have been written “*Vegavis iaa*i, gen. nov., sp. nov.” Without this emendation, only the species name may be available under the ICZN¹.

A formal international code of phylogenetic nomenclature that does not require linnaean ranks is under development³. Even when this is implemented, however, species names should be defined so that they are valid and available under both codes.

1. Polaszek, A., Grubb, P., Groves, C., Ehardt, C. L. & Butynski, T. M. What constitutes a proper description? *Science* **309**, 2165–2166 (2005).
2. International Commission on Zoological Nomenclature. *International Code of Zoological Nomenclature* 4th edn, 18 (International Trust for Zoological Nomenclature, London, 1999).
3. Cantino, P. D. & de Queiroz, K. *Phylocode: a Phylogenetic Code of Biological Nomenclature* Version 3a (www.ohiou.edu/phylocode/) (Ohio University, Athens, USA, 2006).

naturejobs

**THE CAREERS
MAGAZINE FOR
SCIENTISTS**

What do Shakespeare and The Clash have in common? Apart from the fact that they are both English, the bard and the venerable punk group have each posed a well known question. In Shakespeare's *Hamlet*, the troubled Dane asks "To be or not to be?", and in The Clash's 1980s classic, they ponder: "Should I stay or should I go?" Both of these sentiments are relevant to the final instalment of our "Surviving in Science" series. On page 782, we look at how to resolve serious career questions — the kind of decisions that could see you switching from academia to industry, leaving the bench for an off-the-bench position, or even quitting science entirely. We also offer perspectives from people who have grappled with these problems.

To celebrate the close of the series, all of the instalments are available for free on our website (www.nature.com/naturejobs/magazine/editors-choice.html). And, with apologies to The Clash, I offer my take on their song, summarizing the dilemmas faced by scientists, from graduate student to postdoc to junior faculty member:

Doctor, you've gotta let me know/ Should I stay or should I go?
I ask for help, get nothing back/ It's time for you to cut some slack
You say my data do not fit/ Should I stay or should I quit?

This fellowship is number four/ A strategy to keep me poor
E. coli's got it in for me/ Don't know which culture even suits me
My lab book's full to no avail/ Should I stay or should I bail?

I'm simply howling at the moon/ Must get a proper contract soon
Lost out again on tenure track/ Left the lab no turning back
Industry has got the dough/ Should I stay or should I go?

These are the questions. I hope this series has provided a lot of the answers.

Paul Smaglik, *Naturejobs* editor

Editor: Paul Smaglik
Assistant Editor: Gene Russo

European Head Office, London
The Macmillan Building,
4 Crinan Street,
London N1 9XW, UK
Tel: +44 (0) 20 7843 4961
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

European Sales Manager:
Andy Douglas (4975)
e-mail: a.douglas@nature.com
Business Development Manager:
Amelie Pequignot (4974)
e-mail: a.pequignot@nature.com
Natureevents:
Claudia Paulsen Young (+44 (0) 20 7014 4015)
e-mail: c.paulsenyoung@nature.com
France/Switzerland/Belgium:
Muriel Lestrange (4994)

UK/Ireland/Italy/RoW:
Loredana Milanese (4944)
Nils Moeller (4953)
Scandinavia/Spain/Portugal:
Evelina Rubio-Morgan (4973)
Germany/Austria/The Netherlands:
Reya Silao (4970)
Online Job Postings:
Matthew Ward (+44 (0) 20 7014 4059)

European Satellite Office
Germany: Patrick Phelan
Tel: +49 89 54 90 57 11
Fax: +49 89 54 90 57 20
e-mail: p.phelan@nature.com

Advertising Production Manager:
Stephen Russell
To send materials use London address above.
Tel: +44 (0) 20 7843 4816
Fax: +44 (0) 20 7843 4996
e-mail: naturejobs@nature.com

Naturejobs web development: Tom Hancock
Naturejobs online production:
Catherine Alexander

US Head Office, New York
75 Varick Street, 9th Floor,
New York, NY 10013-1917
Tel: +1 800 989 7718
Fax: +1 800 989 7103
e-mail: naturejobs@natureny.com

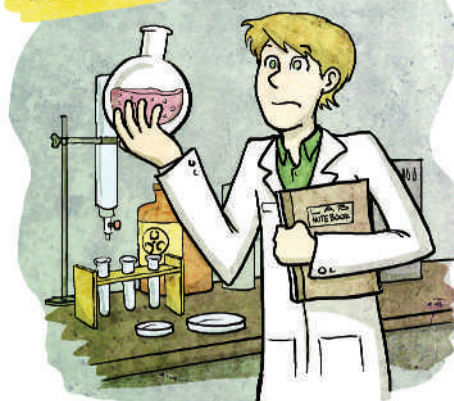
US Sales Manager: Peter Bless

Japan Head Office, Tokyo
Chiyoda Building, 2-37 Ichigayatamachi,
Shinjuku-ku, Tokyo 162-0843
Tel: +81 3 3267 8751
Fax: +81 3 3267 8746

Asia-Pacific Sales Manager:
Ayako Watanabe
e-mail: a.watanabe@natureasia.com



To SCIENCE or not to SCIENCE, that is the question!



Whether it's saner in the mind to suffer the slings and arrows of being a professor...



or to change careers amidst a sea of doubts, and in doing so ...end them?



To shuffle off this campus soil! To teach: no more...



Perchance a dream?

SHOULD I STAY OR SHOULD I GO?

Do you have a nagging feeling that academic research might not be the place for you? Listening to your intuition and trying your hand at new things could place you in your dream career. **Kendall Powell** tests the water.

Four years into her doctoral work in organic chemistry, Sarah Webb began surfing the Internet for potential postdoc positions. Since her first year as an undergraduate, she had always thought she would become a professor at a liberal arts university. That was about to change.

"As I read research descriptions, I had this visceral, gut reaction that said, 'Wow, this is really interesting science, but I don't want to do it,'" Webb recalls. She then had what she describes as her mid-graduate-school crisis and a psychological meltdown. "If you think like an academic, you have your whole life mapped out and then all of a sudden it was 'Oh no!'"

Webb, now a freelance science writer based in Brooklyn, New York, approached her dilemma with a methodical plan of action to find what other careers might suit her. To avoid making a wrong move, career advisers encourage young scientists to make a careful analysis of what they like about science, what their

strengths are, and how they could transfer those strengths to another career track.

Start thinking about your 'plan B' as early as halfway through your doctorate. Even if you think you want to stay in academia, investigate other options. And if you do plan to leave academia, the bench, or even science altogether, you should network and gain experience in the new area before making a switch, career advisers say.

Go with your gut

The academic track is a well-beaten path with a clear set of steps towards a particular destination. It can become comfortable staying on a familiar path, even if your talents and interests no longer match the end goal. "It is easy to get stuck in a rut and end up in the world of someone else's expectations — advisers, colleagues, family," says Webb. "Ultimately, you are the one who has to live with the career expectations you have."

Pay attention to the warning signs, career advisers say. Are you unhappy in the lab because one experiment isn't working, or because a particular colleague is getting on your nerves, or because of trepidation about big-picture career issues? "Do an honest evaluation and be tough on yourself," says Keith Micoli, chair of the board of the US National Postdoctoral Association. And get evaluations from both science and non-science friends and colleagues. Ask them what they see as your professional strengths and weaknesses. Some scientists find they need time away from the research environment to answer these questions.

"Ask yourself what your day job would look like if you could choose it," says Rosana Kapeller, vice-president of research at Renegade Therapeutics in Cambridge, Massachusetts. Do you love working in teams on big projects? Then the pharmaceutical industry might be for you. Do you like reading literature and figuring out where the holes are? Patent law might be for you. Do you love bench work, but hate writing grant applications? You might consider a research associate position.

Micoli says that many young scientists let their fears prevent them from searching out the best career options. It's common to think that one step off the academic path will earn you the label 'not serious about research,' says Micoli, now a research instructor at the University of Alabama, Birmingham. But the sooner you bring up other career interests with your adviser the better, he says.

Taking the blinkers off

How do you go about making a big career transition when you have only been exposed to academia? Michael Alvarez, director of the Stanford School of Medicine's career centre, says that every graduate student should commit to going to at least one seminar or activity per week to explore other career opportunities and make more informed career decisions.

"I'm doing everything I can to eradicate the word 'alternative,'" he says. "There may be 15 alternatives and one of them is academic research. To say another career is lesser is a fallacy. The scientist who does not stay in



LEE FRASER

Philadelphia, Pennsylvania
Lead clinical-development scientist

How did you know academia wasn't for you?

Midway through my master's degree I realized it probably wasn't what I wanted. But I had collected all the badges along the way, so I figured I'd go ahead and get the last badge, my doctorate. But I wasn't completely passionate about the research and I saw that all the people around me who were successful scientists were passionate. I didn't know what the drug industry was all about but I knew there were opportunities there that were different.

Your first industry job was in marketing; what does a science graduate student know about marketing?

I was cycling a lot and had to write a business proposal for a company that we were approaching for sponsorship of our racing team. I found I had a knack for marketing. When you are writing grants or a paper, you always put a certain spin on it, to

make a connection between a disease and the state of your protein. Marketing is just about delivering a set of facts so that people come to the conclusion you want them to — it was a natural extension for me.

Did you make the right choice?

The moment I got my first job and had my own office and sat down at my desk, I knew I had been like a plant in a pot that was too small.

How is working for a large drug company different from academia?

At 2 p.m. on a Friday, I can't shut down and go for a beer because an experiment failed. That freedom is gone. But if you take that 18-hour-day work ethic from graduate school and apply it here, even with weekends off, you'll be really successful. I now work on projects that have budgets greater than the research budget for my old university. But if my kids get sick, my colleagues say: "Go home, we'll take care of it."

academia provides service to the overall well-being of science." This could include expediting drug discovery and approval, teaching at school or university, increasing scientific literacy or improving investment decisions, says Alvarez.

Alvarez also suggests working with a professional career counsellor, consulting books, and doing some rough mental exercises to identify career priorities. In one test, he has the scientist draw a bar graph with three bars, one for geography, one for professional opportunity and one for personal life. The person has 100 units of value to ascribe to the different categories across three different points in time, say at ages 25, 35 and 45.

When trying to decide between two options, make the usual list of pros and cons, but set up a list of categories and weight each category by importance before making your list. Each pro or con item falls into a category and gets assigned a predetermined weight, giving you a more realistic view of which choice aligns with your goals.

Don't make choices based on negatives, says Kapeller.



SUNITA JONES

Manchester, UK
Hospital research facilitator

How did you end up as a research administrator at a hospital?

During the second year of my postdoc, my career plans began to take shape and I realized that bench work was not for me. Interacting with people and helping to get things done was more fulfilling. I had started on the administrative path at Stanford after my postdoc when someone recommended

me to my current boss. I am involved in grant submissions, student and postdoc project reviews, and manuscript preparations.

How was the transition?

Shifting gears and moving away from the bench was definitely the right decision, although I miss my friends and the support network I had established in the United States. I am going through a long period of adjustment and working

hard to establish myself as I am a newcomer to the group. I am enjoying learning the new system.

Is there anything you would have done differently?

I would perhaps start sooner. I think the key was that I did get support when venturing outside my field. My postdoc principal investigator let me explore what I wanted to do beyond my postdoc training and I discussed options with him.

For example, don't choose to move into industry solely because of the downsides to academia, such as writing grants or working long hours. "It's not that the grass is greener on the other side of the fence, but just a different shade of green," she warns. Instead make choices based on what you like about science. But don't be fooled into thinking the next academic stage will be easier, says Micoli. If you are stressed and overwhelmed now, moving up is unlikely to solve your problems.

Find something you're passionate about before fleeing the lab, suggests scientist-turned-artist Tia Vellani. She finished a postdoc in biochemistry at the University of Miami in Florida before deciding to follow her passion for jewellery designing. "It was finally obvious to me that I could never be a good scientist, because I just really didn't want to be," she says.

And finally, if you do decide to leave academia, don't drop a bomb on your adviser by waiting until the last minute to make your plans known. "Leave as many doors open as you can," advises Micoli. "You never know when you might need a recommendation. Be professional."

Landing on your feet

Open and early communication with an adviser may help you find ways to gain experience in a new area. When Webb was searching for a new path, she volunteered to do a few hours a week at a local science museum and enrolled in a science-writing course on her campus. Others have gained insights from volunteering to sit on committees for professional organizations such as the local biotechnology board or even non-science committees, just to build business skills and savvy. Kapeller suggests seeking out a 6–8-week summer internship with a local biotech or drug company (positions that are common, but often unadvertised).

Although some advisers may be dead-set against anything that detracts from time at the bench, most will be reasonable about a request to explore other interests,



MHAIRI DUPRÉ

Oxford, UK
Doctoral candidate in plant science

You left a doctoral programme in Canada and eventually started again in a programme

in Britain. What made you decide to come back to academic science?

When I left Canada, I came back to my village of Ballachulish, Scotland, and began working as a waitress. It was a good break, to be around people just doing the job for the money, to try to figure out what I really want to do. I realized that I do like science and that I have not really wanted to get up and work in a bank or be a lawyer every day. Also, I went to different lectures on history and philosophy to see if I'd be good at those, but when I was reading newspapers, I would see the science stories and think, "Oh, that's

really interesting." To be able to change people's lives through science, to discover something new — no other job can do that.

Would you recommend a break from science for others who are undecided?

For sure. Otherwise, I just would have rushed into another decision. It gave me time to have no pressure and find out what really excites me.

Would you do anything differently?

If I was beginning a PhD again, I would try to work in the lab a bit before starting. Ask yourself, can I do my PhD in this lab, can I get along with this supervisor, can I do the project I want to do? Also, don't be scared to say "This isn't working" and cut your losses.

says Micoli. Explain that you would like to take on more teaching duties, offer to help the technology-transfer office with a patent application, or suggest an industry internship that will lead to a collaboration. Webb negotiated with her adviser to have two months away from the lab to work half-time on writing her thesis and half-time building her science journalism portfolio.

Most importantly, find people who are already doing the job you want to do and talk to them about their own transition. See if you can visit them at their workplace or shadow them for a day. If possible, find someone who has made exactly the same transition that you are contemplating.

For those pondering a switch to industry, Kapeller strongly advises doing an academic postdoc before making the jump. Not only will it let you step on to the corporate ladder on a higher rung, she says, but it will confirm your ability to work and publish independently more effectively than the doctorate alone or an industry postdoc would. If you know you will be moving to industry, she suggests choosing a postdoc with a focus on animal models,



KATY HINMAN

Atlanta, Georgia
Executive director, Georgia Interfaith Power and Light

What is Interfaith Power and Light?

It is a non-profit organization that works with congregations and faith communities on environmental issues, with a particular focus on energy. We do a lot of practical education about energy efficiency. Right now, we are making our compact fluorescent light-bulb kits to distribute during the Chanukah and Christmas holidays.

How did you end up working at the interface of science and religion?

After I finished my doctorate, I became really interested in why people weren't more

interested in conservation — particularly in faith communities. They should be thinking about environmental stewardship just like the other ways they think about stewardship. I had been active in my church throughout graduate school and it really seemed to me that, in general, the relationship between science and religion was one people don't talk about.

Why did you decide to go to seminary on top of your PhD in ecology?

My thought was, "Well, of course churches should be involved in conservation", but that argument doesn't fly with most people. It got me thinking about making a theological case

for them to get involved. My scientific and seminary training helps me cross that gap. I have credibility on both sides.

Would you recommend working for a non-profit organization to others?

Yes, these organizations need people who are good at science — people who know how to interpret it and translate it.

Do you miss research?

I really miss doing fieldwork and talking with other scientists. So I go to the North American Symposium on Bat Research just for my own intellectual stimulation.



PETE BERQUIST

Williamsburg, Virginia

Research assistant, National Park Service employee, adjunct professor

You were encouraged by your master's supervisor to stay on and get a PhD in geology. What made you decide against it?

Overall, graduate school was a great experience. I loved doing research and teaching, but both of these took so much time. One conversation I had with a friend stuck with me. He asked: "Are you working too hard?" And I said: "I don't know, how do you tell?" His response was telling, he said: "Well, are you having fun?" I felt as though there were other parts of life that I was missing.

So what have you found?

I went to Maine and worked for a non-profit advocacy group in Acadia National Park. I was outside hiking all day, so it was hard to call it a job. I then interned with the National Park Service, teaching at a residential science camp for middle-school students and teachers. I came back to work with my undergraduate adviser as a research assistant doing geologic mapping. He jokes that I'm doing a "post-master's" instead of a postdoc. And I've been a visiting professor, teaching geology at the College of the Atlantic, a very small liberal arts college in Bar Harbor, Maine.

What have you learned from these experiences?

I discovered that I want to pass on and share what I've learned with people who may not have had a lot of geology or Earth science. Interacting with a more diverse group of people than in academia, it is satisfying to introduce a geological perspective that they might not get anywhere else.

What advice do you have for others searching out their own career paths?

Be open. There's a lot of room in there to open your eyes to new things. Every time I've done that, I've made new contacts.

pharmacology or imaging that would be applicable in a corporate setting.

Those who have left science suggest making sure that you can live with the prospect of never being a scientist again. Being away from the swift changes in the literature and technologies of specialized fields for even a few years can make returning an uphill battle.

And finally, maybe you could learn from Katy Hinman, the executive director of Georgia Interfaith Power and Light, a non-profit organization in Atlanta that counsels religious communities about environmental stewardship. Her job certainly never appeared in any 'alternative careers' books or panels. But she identified two things that were important to her — conservation and her faith — and followed where they led after her PhD in ecology and evolution from the State University of New York in Stony Brook, even though it meant going to seminary.

"People get into a kind of trap, thinking that if a job doesn't require a PhD in its description, then they are underemployed," says Hinman. "If you are doing something you love and are good at it, then you are not underemployed."

Kendall Powell is a freelance writer based in Broomfield, Colorado.

Web links and further reading

Stanford SOM Career Center profiles

▶ <http://med.stanford.edu/careercenter/spotlight>

US National Postdoctoral Association's career development resources

▶ www.nationalpostdoc.org/site/c.eoJMIWOBIrH/b.1389993/k.B38F/Career.htm

Tia Vellani's site

▶ www.artistbynight.com

Sarah Webb's site

▶ www.sarahannewebb.com

Bolles, R. N. *What Color is Your Parachute?* (Ten Speed Press, 2006).

Covey, S. *The 7 Habits of Highly Effective People* (Free Press, 2004).

Kreeger, K. Y. *Guide to Nontraditional Careers in Science* (Taylor & Francis, 1998).

Lloyd, C. *Creating a Life Worth Living* (Harper, 1997).

Robbins-Roth, C. *Alternative Careers in Science: Leaving the Ivory Tower* (Academic Press, 2005).



STACEY IVANCHUK

Toronto, Ontario, Canada

Training to be a patent agent

Why did you leave academic science for patent work?

Seven years after starting my doctorate, I was struggling to finish it. I had two potential manuscripts, but nothing that was going to point me in the direction of being a professor. I have a lot of friends who are lawyers, familiar with intellectual-property law, and they suggested that maybe I was burnt out on the lab and needed a new perspective. They said, "It's still science, but coming at it from a different angle."

What advice do you have for others?

You need to get out of the lab and see

what else there is. It was only after I left that I started to learn about networking — that getting a job is sometimes about timing and sometimes about who you know. Be more proactive and think, "This is what I want to do: who do I need to know to make this happen?"

What are the pros and cons of working for a law firm?

Suddenly, it's no longer the flexibility of the lab. You have a deadline, sometimes before the day is over — you must get back to a litigator by 5 p.m. That part takes some getting used to. As for the pro side, I get exposure to so many different inventions in the world of molecular biology.



Awakening the genius within

Revolution in the head.

Daniel Gregory

"From ordinary to extraordinary, this machine will make you Shakespeare! Plato! Einstein! Buy it now and awake the genius within! The Inspiron-4000 thinks so you don't have to."

The holographic TV blurted the advert out at a speed too quick to understand at first, but moments later comprehension dawned.

So Martin Williams sat at his desk and thought: a bare featureless landscape lay before him... a daunting nothingness, the blank page stared back at him. Uninspired. Then he realized what he'd heard.

As a would-be inventor, Martin understood that the key to success was inspiration. None was forthcoming. Time and again he sat down at his desk, only to shrink away from the blank page before him. His life had been reduced to begging and borrowing. He'd drunk and done drugs and turned his modest, beautiful home into a dirty, smelly pigsty. Papers, food, cans of beer accumulated as remnants of his torn-up life. The pieces were everywhere.

He needed ideas.

Martin pressed the buttons on his TV remote and purchased the Inspiron-4000. The Inspiron, which attached directly to the cerebral cortex, was well-known for providing inspiration to many thousands of budding thinkers — inventors, painters, scientists. And the 4000 model was the second in a series, an update of the 3000 model — it should be even better. (The Inspiron-3000 had apparently inspired its original inventors to greater heights — a self-sustaining system of continual and exponential improvement. Each machine gave the inspiration to improve itself.)

A light flashed. Through the magic of teleportation, the Inspiron-4000 materialized in Martin's lounge. He sparkled to life, his senses awake as he tore the packaging apart to reveal the gem inside. Ears and eyes on overdrive, he heard the gentle hum of a far-off space-launch and his eyes recorded a gliding hover-car pass by.

The Inspiron-4000 gleamed in the sunlight. Its surface was metallic and shone silver in the bright sun. It was perfectly

shaped. He fitted it on smoothly. The world faded.

Martin drifted, agile and omnipotent, across the world. Visions appeared like dreams, tangible yet slippery, he gripped at them with his mind. He felt split between worlds — the visions and reality. Snatching at paper he wrote and wrote — great, great works. He grabbed at inspiration, plucking ideas from the ether.

Martin ate and drank nothing. Days passed. He was obsessed and only quietly, from a distance, it seemed, he heard a banging.

"Open up! It's me, Jim."

No one answered. "Open up!"

"Coming...coming," said Martin. "Yeah,

capable of! I'll be a new Shakespeare! Plato! Einstein!"

Jim laughed at the ridiculousness of it, but he halted as he saw the look in Martin's eyes. For a moment, he was taken aback. "No — don't use it. Take it off. It's dangerous. Take it off," he repeated. His voice rang with doubt.

"But look — LOOK!" said Martin. "All these works: beautiful, beautiful works! The Inspiron's made something of me — don't you get it? — it's made me a genius. Before I was no one. Now I'm a god! Worship me!"

His dark eyes welled up with tears and a desperate frown spread across his face. He seemed split between two worlds: magnifi-

cent papers were strewn across a dingy, messy room. He had done all these things. He was, as he had said, a genius — a god — yet he seemed so animal-like: a sad shadow come full circle from failure to success, and now he was sliding back. He seemed alone. He seemed to teeter on a knife-edge.

"I'm tired of losing," he said.

"Take the damn thing off!" said Jim.

"No!"

"Take it off! For God's sake, save yourself! Take it off!"

"No."

"Please."

Martin wavered. He struggled, unsure of himself. "Why don't you try it for yourself? You want to, don't you? You'll understand then." A serene, creepy voice. Jim was tempted. He wavered. "I've made another one," said Martin. "I've got two."

A pause. The world stood still.

Jim placed the Inspiron slowly on his head. He saw the visions, the ideas, the possibilities.

"Take it off," said Martin.

"You take it off!"

"No. Maybe — wait!"

In the distance there was the gentle hum of a far-off space-launch. A hover-car glided past. And here had seemed something so hopeful. Teetering on the knife-edge, the humans struggled against the Inspiron and the talent that it could give them. ■

Daniel Gregory lives on Anglesey, North Wales, UK. He can be contacted at dangreg999@aol.com.



JACEY